Gauge Theories: a Case Study of how Mathematics Relates to the World

A thesis presented

by

Antigoni Nounou

to

University of London

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Philosophy

London School of Economics and Political Sciences

University of London

London

April 2002

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The goal of this thesis is to investigate the relation between mathematics and physics and the role this relation plays in what physics does best, that is in scientific explanations. The case of gauge theories, which are highly mathematical, is used as an extended case study of how mathematics relates to physics and to the world and these relations are examined from both a historical and a philosophical perspective.

Gauge theories originated from an idea of Weyl which turned out to be wrong, or in other words, empirically inadequate. That original idea underwent a dramatic metamorphosis that turned the awkward caterpillar into a beautiful butterfly called gauge theories, which were very successful and dominated theoretical physics during the second half of the twentieth century. The only leftover from Weyl's *faux pas* was the very name of the theories and the question how it is possible for something as wrong as his original idea to result in a theory so relevant to the world. We argue that it is thanks to a very dynamic and dialectic relation between mathematicians and physicists, both theoretical and experimental, that the resulting theory turned out to be so successful.

From a more philosophical perspective, we take the view that the relation between mathematics and physics has a structuralist character, in general, and we recognize that what we call ambiguity of representation of the third type lies at the heart of gauge theories. Our claim is that it is precisely this type of ambiguity of representation and the non-physical entities that it inevitably introduces which ex-

plain the physical facts. However, the non-physical entities should be attributed a non-causal status in order to provide valid and legitimate scientific explanations. The fibre bundles formulation of gauge theories is considered to be their unique formulation that allows for this shift and the Aharonov-Bohm effect which is examined within the fibre bundle context provides a narrower yet very fruitful case study.

Acknowledgments

I would like to thank the Greek State Scholarship Foundation for their generous support over the first three and a half years of this Ph.D. and the LSE Department of Philosophy, Logic and Scientific Method for the Popper scholarship and for the financial support they offered me for a year. Also, I would like to thank all the people who helped me in completing this thesis. Craig Calender Nancy Cartwright, Jordi Cat, Carl Hoefer, Jeff Ketland and Stathis Psillos contributed, one way or another, to the opening of my philosophical horizons. From Imperial College, Chris Isham and Kelly Stelle patiently explained to me again and again difficult technical issues and guided me through gauge theories, gravity and fibre bundles. I am most grateful to Michael Redhead, D.K. and my mother because without their support on all levels this thesis would have never been completed. Finally, I am indebted to my two examiners, Dr. Chang and Professor Kilmister for their valuable comments that helped me improve the thesis.

Contents

[1] The title of this section is borrowed from a phrase that can be found in O'Raifeartaigh's *The Dawning of Gauge Theory*, p.7. O'Raifeartaigh's book is highly recommended as a wonderful resource for more precise and complete historical detail. For a standard physics introduction to this material reference may be made to Aitchison & Hey's *Gauge Theories in Particle Physics*.

Introduction

The motivation behind this thesis has been my wonderment while I was still doing my degree in physics and I was first introduced to the notion of covariant derivative in an undergraduate course on general relativity. The idea that spacetime itself is modified when there are sources of gravitational field present I found extraordinarily illuminating. A year later, while I was working on a project on elementary particles, I was introduced to gauge theories and I was surprised when I saw that the notion of covariant derivative was so far-reaching that it appeared there as well. Since then, I have been trying to understand what was the connection between gravity, on one hand, and quantum field theory, on the other, that appeared in the form of the covariant derivative that was present in, what I perceived then as, two theories. My curiosity about and my reasons for being attracted to theories that involve covariant derivatives, along with the conclusions I have reached are presented in the thesis that follows.

To the amazement of many who are interested in history of ideas, there appear to be many incidents in the history of physics where the mathematics that was needed for the accurate formulation of a physical theory was already there when physicists needed it. Something like that seemed to have happened in the case of the physical gauge theories and the mathematical fibre bundles, because although there was no apparent interaction between the two communities, when gauge theories were mature enough to make use of the fibre bundles formalism, the formalism was already there, mature and ready. But those who would like to tell a story like this actually overlook that at the heart of both theories

lies the same idea, an idea of Herman Weyl. His original idea, dated back to 1918, did not apply to the world, as Einstein pointed out immediately after Weyl formulated it. Yet despite the Einstein's criticism, which he expressed in a series of letters he exchanged with Weyl and which we examine in the first chapter of this thesis, the idea was adopted by others and hammered into something different that maintained the original spirit though. The aim of this idea was to bring together electromagnetism and general relativity, the two then known fundamental theories of nature, something which was eventually achieved very successfully and very fruitfully when Weyl's original scale factor became a phase factor. Both the original and the transformed ideas were related to symmetry transformations and parallel transport. These were ideas that were adopted and developed by mathematicians as well, who delivered the theory of fibre bundles within three decades, while the physicists had to spend five or six decades at a much slower pace before their theory was able to meet with the fibre bundles. These are the main ideas about the dialectic relation between physics and mathematics explored and analyzed in the first chapter.

From the second chapter onwards the thesis takes a different turn and investigates the relation between physics and mathematics from a philosophical perspective. In that chapter we are asking whether we can do science without mathematics, as Field claims. We argue that at least in the case of gauge theories this is not possible and since we answer in the negative, we take on board Redhead's structuralist ideas. According to these ideas, mathematical structures relate to physical systems through mappings and involve what Redhead calls ambiguity of representation, which comes in three types. This approach fits our case study very well for two reasons. The first is that gauge theories, especially when they are

formulated using fibre bundles, deal with nothing other than mappings, between the space-time manifold -or the real world, we might say- and the bundles, or between the bundles themselves. Our second reason for favouring this approach is that the ambiguity of representation of the third type always involves surplus structure and if gauge theories are known for something this is their own surplus structure, namely the gauge potentials themselves. So if mathematics relates to the world with mappings we'd better have a look at those mappings and if the surplus structure has something to say about how things are and how they behave 'down' in the structure and in the world, we'd better find out what this is.

Gauge theories may be formulated in three ways: as constrained Hamiltonian systems, as Yang-Mills theories and using fibre bundles. The three formalisms are intertranslatable to each other, yet, in our view, it is in the last two formulations that we may see more clearly what is the role that the various entities of the theory play. In the third chapter we discuss them all but we emphasize on the last two because from the second we can see easily how the interaction terms arise from symmetry considerations, while from the third we get the most general picture of the entire theory, with its mappings and its surplus structure.

The surplus structure and its own role in both 'controlling' the physical system and providing scientific explanations is the main topic of the fourth chapter. We dedicate quite a long discussion on the Aharonov-Bohm effect in it because electromagnetism is the simplest of the gauge theories and because the effect itself provided the first inkling that objects of the surplus structure may be something more than just disposable mathematical artifacts that only simplify calculations. If it turned out that the objects of the surplus structure are

indeed more than mere devices, then their status in the theory and their explanatory and predictive roles would reveal a lot about the relation between the mathematical and the physical.

Ever since it was discovered, the effect required an explanation that is valid and meaningful. From the very existence of the effect, it became clear that the electromagnetic field does not suffice to give a local and causal explanation, hence some other entity should provide that kind of service. The first idea was that since the effect is described by the gauge potential, which is present in all the regions concerned, the gauge potential itself might do the trick. However, there are difficulties in assigning to that field any local and/or real character; hence something else might be necessary, either another entity that would be able to play the local-causal role or a different interpretation of the gauge potential and its role. We argue that the right answer may be found in the second suggestion because if we consider that the surplus structure 'controls' the physical only in an informative and descriptive, rather than causal, sense and that what it describes is topological properties, then we get holistic topological explanations. Topological explanations are a distinctive kind of explanation and so far as classical field theory is concerned they are not good explanations, not even as approximations. Things change, however, in the case of relativistic quantum field theory where the holistic explanations are valid and far reaching.

The fact that this adjustment in the way we interpret the status of the gauge potential provides a valid explanation of an admittedly significant effect reveals one more aspect of the dialectic relation between physics and mathematics, we argue. Theories are used for explaining certain phenomena and predicting other, yet undiscovered, phenomena even

though interpretational issues may still be subject to debate and revision. The explanatory power of these theories increases as further evidence comes -or is expected to come- into light and guides us towards further modification of our views and of our interpretations.

Chapter 1
Some History

A gauge, according to the dictionary, is "a standard measurement, dimension, capacity or quantity; a standard or means for assessing". It is also "any of various devices used to check for conformity with a standard measurement". The term gauge as a noun seems to have been used in physics in three different contexts. We measure pressure using a pressure gauge; Maxwell's equations of electromagnetism are known to be invariant under a symmetry transformation called a gauge transformation; and finally, we describe the fundamental forces of nature using gauge theories, gauge symmetries and gauge fields. The common denominator in all the uses of the word 'gauge' above is that there is an element of arbitrariness of choice involved. A standard measurement, for example, is standard because we have chosen it to be so, but this choice is arbitrary. The way we have calibrated the pressure gauge is also arbitrary, in the sense that we could choose any other scale. In classical electromagnetism, since the gauge transformation is a symmetry transformation, that is to say leaves the equations unchanged, the choice of a specific gauge is also arbitrary and a matter of convention. Finally, in the case of fundamental interactions there is also some arbitrariness involved, which is related to the fact that these interactions are described using gauge symmetries, but this arbitrariness will be examined in detail later in this thesis, since it is in the context of gauge theories that the term gauge has been used the most extensively in modern physics.

In this chapter we will delineate the surfacing of gauge theories in physics and we will focus on the dispute that the first attempt to give a unified theory of gravitational and electromagnetic interactions produced between Weyl and Einstein, while at the same time we will try to specify the meaning of the term 'gauge' in the various periods that it was used. In the last section we will also try to shed light on how a whole new branch of differential geometry, which accommodates gauge theories in a most comprehensive way, developed almost in parallel with them.

1.1 The Quest for the holy Grail of a Unified Theory

It all started with electromagnetism, general relativity and Weyl's quest for the holy grail of a unified theory of the two. Or rather the quest for an "archetype geometry" , as Ryckman (2001) calls it, that could accommodate all possibilities of physics. During the time Weyl was working on his geometry, there were known only two interactions which were considered to be elementary: the electromagnetic and the gravitational. Hence, these two were the possibilities of physics that should be described by his geometry. Electromagnetism was known to be a gauge invariant theory since its discovery, but this property of the theory was not given any geometrical significance or physical interpretation. Instead, the formulation of Maxwell's equations in terms of the gauge field, then known as the vector potential, was used merely because it made certain calculations easier. But Weyl's quest for a unified geometry that should be able to account for both the gravitational and the electromagnetic interactions led to a theory in which a geometrical significance was attributed to that field.

Weyl completed his endeavour by 1918, and the main idea in it was that since Riemannian geometry described successfully the gravitational field, maybe a more general affine geometry[2] would describe both gravitation and electromagnetism in a unified way. The question that had to be answered was, of course, which affine geometry. Weyl begins his 1918 paper as follows.

"According to Riemann, geometry is based on the following two facts:

1. *Space is a three dimensional continuum*, the manifold of its points is therefore represented in a smooth manner by the values of the three coordinates x_1, x_2, x_3.

2. (*Pythagorean Theorem*) The square of the distance between two infinitesimally separated points

$$P = (x_1, x_2, x_3) \text{ and } P' = (x_1 + dx_1, x_2 + dx_2, x_3 + dx_3) \quad (1)$$

is (in any coordinate system) a quadratic form in the relative coordinates dx_i:

$$ds^2 = \sum_{ik} g_{ik} dx_i dx_k \qquad (g_{ik} = g_{ki}) \quad (2)$$

We express the second fact briefly by saying: the space is a *metrical* continuum. In the spirit of modern local physics we take the Pythagorean theorem to be strictly valid only in the infinitesimal limit.

Special relativity leads to the insight that time should be included as a fourth coordinate x_0 on the same footing as the three space-coordinates, and thus the stage for physical events, *the world*, is a *four-dimensional, metrical continuum*. The quadratic form (2) that defines the world-geometry is not positive-definite as in the case of three-dimensional geometry, but it has a positive index-3. Riemann already expressed the idea that the metric should be regarded as something physically meaningful since it manifests itself as an effective force for material bodies, in centrifugal forces for example, and that one should therefore take into account that it interacts with matter; whereas previously all geometers and philosophers believed that the metric was an intrinsic property of the space, independent of the matter contained within it. It was on the basis of that idea, for which the possibility of fulfillment was not available to Riemann, that in our time Einstein (independently from Riemann) erected the grandiose structure of general relativity. According to Einstein the phenomena of gravitation can be attributed to the world-metric, and the laws through which matter and metric interact are nothing but the laws of gravitation; the g_{ik} in (2) are the components of the gravitational potential.-Whereas the gravitational potentials are the components of an invariant quadratic differential form, *electromagnetic phenomena* are controlled by a four-potential, whose components ϕ_i are components of an invariant linear dif-

[2] Affine meaning length preserving.

ferential form $\sum_i \phi_i dx_i$. However, both phenomena, gravitation and electricity, have remained completely isolated from one another up to now"[3].

Ryckman (2001) has argued that although phenomenological evidence was important, most crucial for Weyl were sensation and intuition. Truth for him was identified with the experience of truth, which did not have to rely necessarily on perception. Hence, despite the fact that it had not been observed, Weyl considered as a leading principle the a priori relativity of length and claimed that "[*a*] *true infinitesimal geometry should, however, recognize only a principle of transferring the magnitude of a vector to an infinitesimally close point* and then, on transfer to an arbitrarily distant point, the integrability of the magnitude of a vector is no more to be expected than the integrability of its direction. On the removal of this inconsistency there appears a geometry that, surprisingly, when applied to the world, *explains not only the gravitational phenomena but also the electrical.* According to the resultant theory, both spring from the same source, indeed *in general one cannot separate gravitation and electromagnetism in an arbitrary manner.* In this theory *all physical quantities have a world-geometrical meaning; the action appears from the beginning as a pure number; it leads to an essentially unique universal law; it even allows us to understand in a certain sense why the world is four-dimensional*"[4]. This requirement for relativity of length involves the arbitrariness of choice of what one should call a unit of length -hence the term gauge becomes relevant- and gives an affine geometry which differs from the Riemannian in the following sense. While in Riemannian geometry the inner product between vectors is invariant, in Weyl's affine and metrical vector-space the invariant scalar product

[3] Weyl, 1918.

[4] Weyl, 1918.

of two vectors defined at a point P

$$\chi \cdot \eta = \eta \cdot \chi = \sum_{ik} g_{ik} \chi^i \eta^k$$

"*is determined only up to an arbitrary positive proportionality-factor*"[5]. Hence, the metric at P determines not the components g_{ik} themselves, but the ratios of the components. This entails that at each point of a manifold one has the freedom to choose the coordinate system as well as the proportionality factor of g_{ik}. If one requires that every formula of the theory is invariant under both arbitrary smooth coordinate transformations and the transformation $g_{ik} \rightarrow \lambda g_{ik}$ and also one defines the parallel transfer of a vector at P_1 to a neighboring point at P_2 by the following axioms:

1. the parallel transfer of the vectors at P_1 to vectors at P_2 defines a similarity map

2. if P_1 and P_2 are two neighboring points to P and if the infinitesimal vectors $\overrightarrow{PP_2}$ and $\overrightarrow{PP_1}$become $\overrightarrow{P_1P_{12}}$ and $\overrightarrow{P_2P_{21}}$, on parallel-transfer to P_2 an P_1 respectively, then P_{12} and P_{21} coincide (commutativity)

then for a vector $\xi^i \rightarrow \xi^i + d\xi^i$ one gets

$$d\xi^i = -\sum_r d\gamma_r^i \xi^r$$

The second axiom requires that the $d\gamma_r^i$ are linear differential forms

$$d\gamma_r^i = \sum_s \Gamma_{rs}^i dx_s$$

where

$$\Gamma_{sr}^i = \Gamma_{rs}^i$$

---

[5] Weyl, 1918.

If two vectors are parallel transferred, the part of axiom 1 that goes beyond affinity to include similarity requires that the scalar product of the original vectors ξ^i and η^i is proportional to the scalar product of the transferred vectors $\xi^i + d\xi^i$, $\eta^i + d\eta^i$. If the proportionality factor is $(1 + d\phi)$ we get

$$\sum_{ik} (g_{ik} + dg_{ik})(\xi^i + d\xi^i)(\eta^k + d\eta^k) \propto \sum_{ik} g_{ik}\xi^i\eta^k$$

$$\sum_{ik} (g_{ik} + dg_{ik})(\xi^i + d\xi^i)(\eta^k + d\eta^k) = (1 + d\phi) \sum_{ik} g_{ik}\xi^i\eta^k$$

and finally we have

$$dg_{ik} - (d\gamma_{ki} + d\gamma_{ik}) = g_{ik}d\phi \quad (6)$$

From this expression follows that $d\phi$ is a differential form:

$$d\phi = \sum_i \phi_i dx_i \quad (7)$$

When ϕ is known, then the quantities Γ are determined by the equation

$$\Gamma_{i,kr} + \Gamma_{k,ir} = \frac{\partial g_{ik}}{\partial x_r} - g_{ik}\phi_r.$$

Hence, "*the metrical connection of the space depends not only on the quadratic form (2) but on the linear form (7)*"[6]. So, as a result of the additional requirement for similarity -that goes beyond affinity- the quantities Γ depend not only on the derivatives of the metric, but also on a vector field ϕ. The physical significance that can be attributed to these quantities arises then from the following considerations.

If, first of all, we consider a transformation of the metric $g_{ik} \rightarrow \lambda g_{ik}$ and keep the coordinates the same, the $d\gamma^i_r$ remain the same, $d\gamma_{ir} \rightarrow \lambda d\gamma_{ir}$, and $dg_{ik} \rightarrow \lambda dg_{ik} + g_{ik}d\lambda$.

---

[6] Weyl, 1918.

Varying equation (6) then we get

$$d\phi + \frac{d\lambda}{\lambda} = d\phi + d(\ln \lambda).$$

Hence, for the linear form $\phi_i dx_i$ the arbitrariness takes the form of an *additive total differential* rather than a proportionality factor that would be determined by a choice of scale. This tells us that the forms

$$g_{ik}dx_i dx_k \quad and \quad \phi_i dx_i$$

in Weyl's geometry are equivalent to the forms

$$\lambda g_{ik}dx_i dx_k \quad and \quad d_i\phi_i + d(\ln \lambda)$$

respectively. The quantity that remains invariant under the scalar factor transformation is therefore the antisymmetric tensor

$$F_{ik} = \frac{\partial \phi_i}{\partial x_k} - \frac{\partial \phi_k}{\partial x_i}.$$

This antisymmetric tensor satisfies the first set of Maxwell equations and hence it could be identified with the electromagnetic field. When the coordinates do not undergo a transformation and the parallel transfer of a vector does not depend on its path, then g_{ik} can be chosen so that ϕ_i vanishes. In this case, Γ^i_{rs} is the Christoffel 3-index symbol. As Weyl points out, "once the concept of parallel-transfer is defined the geometry and tensor calculus is easily deduced"[7]. Here we will not take the trouble to show how this is done[8], but we feel obliged to mention how both gravity and electromagnetism arise in the same way from this one geometry.

---

[7] Weyl, 1918.

[8] The reader may look at Weyl's original paper.

Assuming that "the whole set of natural laws is based on a definite integral-invariant, the action", Weyl writes an action of the form

$$\int W d\omega = \int R^i_{jkl} R^{jkl}_i d\omega = \int W\sqrt{g}dx = \int \mathcal{W}dx$$

where $R^i_{jkl} = P^i_{jkl} - \frac{1}{2}\delta^i_j F_{kl}$ are the components of the analogue of the Riemann curvature tensor where $P^i_{jkl} = 0$ in the absence of gravitational field, while $F_{kl} = 0$ in the absence of electromagnetic field. In general, $W = 0$ only in the Euclidean space. "*The actual world*" Weyl writes, "*is selected from the class of all possible worlds by the fact that the Action is extremal in every region* with respect to the variations of the action which vanish on the boundary of that region"[9]. Varying this action, therefore, and requiring the variation to vanish on the boundary we have

$$\delta \int \mathcal{W}dx = \int (\mathcal{W}^{ik}\delta g_{ik} - \mathbf{w}^i \delta\phi_i)dx$$

from which we get the field equations

$$\mathcal{W}^{ik} = 0 \quad \text{and} \quad \mathbf{w}^i = 0$$

which are the equations for the gravitational and the electromagnetic field respectively. Five out of these equations may be obtained if one requires invariance of the action under infinitesimal coordinate transformations and under scale transformation. These, obviously, correspond to invariance properties of the action and hence are dubbed superfluous by Weyl. Yet, these equations correspond to the conservation law of the electromagnetic charge and the energy-momentum conservation equations.

---

[9] Ibid.

1.2 The Weyl-Einstein Debate

In 1918, Weyl was working on his geometry, but as he foresaw that the "calculational execution of the theory"[10] would take him quite sometime before it was completed, he decided to publish a report on its foundations beforehand. For that reason he contacted Einstein with the request that he might present it to the Berlin Academy. Einstein responded swiftly to Weyl's request, to whom he wrote in the following day that his work was "a first-class stroke of genius"[11]. But Einstein only took nine days to formulate what he called his "measuring-rod objection"[12]. Einstein's main concern was agreement with reality and on the 15th of April 1918 he was able to assert confidently that "[a]s pretty as your idea is, I must frankly say that in my opinion it is out of the question that the theory corresponded to nature"[13]. The reason for Einstein's objection lies at the heart of Weyl's geometry, namely his assumption that the action remains invariant under a re-scaling of the metric. Such a rescaling, as we have seen, renders ds and λds equivalent. But for Einstein, "ds itself has real meaning"[14] in the sense that if two rigid rods of equal length travelling from point P, where they were at relative rest, to point P', were they are at relative rest again, their relative lengths must be equal. But with Weyl's λ-factor that is arbitrary, the ratio of the two lengths would depend on the paths the two rods follow -or on the arbitrary scale related to those paths. Einstein's original argument was about clocks and is the following.

---

10 Letter from Elmshorn, 5th April 1918.

11 Letter to Weyl, 6 April 1918.

12 Ibid.

13 Letter to Weyl, 15 April 1918.

14 Ibid.

"Imagine two clocks running equally fast at rest relative to each other. If they are separated from each other, moved in any way you liked and then brought together again, they will again run equally (fast), i.e. their relative rates do not depend on their prehistories.

Imagine two points $P_1\&P_2$ that can be connected by a timelike line. The timelike elements ds_1 and ds_2 linked to $P_1\&P_2$ can then be connected by a number of timelike lines upon which they are lying. Clocks travelling along these lines give a fixed relation $ds_1 : ds_2$ independent of which connecting line is chosen. If the relation between ds and the measuring -rod and clock measurements is dropped, the theory of relativity loses its empirical basis altogether"[15].

Apparently, Einstein was not the only one to object to Weyl's idea. In his 19th of April 1918 letter to Weyl, Einstein reports that when he presented the paper on the 11th of April, Nernst "stood up and protested against acceptance of the paper without further comment; he demanded that I at least attach a note in which I describe my different standpoint. Planck then suggested I consider the matter for a week and then submit the paper again, with or without comment, as I consider appropriate". Finally, Einstein suggested that Weyl should include his objection as a postscript and in the same letter he phrases this as follows.

"If light rays were the only means of establishing empirically the metric conditions in the vicinity of a space-time point, a factor would indeed remain undefined in the distance ds (as well as in the $g_{\mu\nu}$'s). This indefiniteness would not exist, however, if the measurement results gained from (infinitesimal) rigid bodies (measuring rods) and clocks are used in the

---

[15] Ibid.

definition of ds. A timelike ds can then be measured directly through a standard clock whose world line contains ds.

Such a definition for the elementary distance ds would only become illusory if the concepts 'standard measuring rod' and 'standard clock' were based on a patently false assumption; this would be the case if the length of a standard measuring rod (or the rate of a standard clock) depended on its prehistory. If this really were the case in nature, then no chemical elements with spectral lines of a specific frequency could exist, but rather the relative frequencies of two (spatially adjacent) atoms of the same sort would, in general, have to differ. As this is not the case, the fundamental hypothesis of the theory unfortunately seems to me not acceptable, the profundity and boldness of which must nevertheless instill admiration in every reader"[16].

This note was, in fact, included as a postscript to Weyl's paper when it was published by the Academy and it was followed by Weyl's reply, who did not seem to agree with Einstein's point after all. Weyl disagreed because he considered that rods and clocks may undergo changes as they move into electromagnetic and gravitational fields, hence they do not constitute appropriate experimental evidence that there is no place for an arbitrary scale factor in the theory. Light signals, on the other hand, determine the absolute values of the metric, yet he considered it an assumption that ds was normalized the way it was, i.e. so that the scale factor was equal to unit in the absence of electromagnetic field or in the presence of a static one. This assumption, Weyl believed, was in need for both an explicit dynamical calculation, in Einstein's theory as well as his, and experimental

[16] Letter to Weyl, 19th April 1918.

verification. The experimental verification would be the red shift of atomic spectral lines in the neighborhood of large masses. A very interesting thing about Weyl was his persistence in his theory even after he found out that the eagerly awaited red shift was not observed, as he reported to Einstein in a letter written on the 18th of September 1918. As a matter of fact, his response to Einstein's objection was posted on November the 16th 1918, two months after he found out that there was no red shift, and yet in it he still insisted on his position and on the need for further experimental verification.

But what was it that gave Weyl the courage to defend his position? He himself admitted in his 10th of December 1918 letter to Einstein that he was "now in a really difficult position; through my upbringing so conciliatory by nature that I am almost incapable of discussion, I must now fight on all fronts". Weyl felt he had to fight on both the mathematics' and the physics' fronts since his analysis and original idea were attacked by the mathematicians, while the physical implications of his geometry raised a debate within -or attack from- the physicists community. To our knowledge, the best explanation was given by T. Ryckman (2001), who claims that Weyl's persistence in his idea arose from a deep philosophical and metaphysical guiding principle rather than 'reality' and physical intuition. What we may get out of this all -apart from the very fact that this mistaken first attempt became Ariadne's thread that lead to gauge theories as we know them today- is that Weyl realized that there might be a way of unifying the electromagnetic and the gravitational interactions rather accidentally. As he himself confessed[17] to Einstein, he ended up introducing the linear differential form along the quadratic one because he wanted to re-

[17] Letter to Einstein, 10th December 1918.

move what he called an "inconsequence"[18]. In his own words, "[i]ncidentally, you must not believe that I came via physics to introducing the linear differential form $d\phi$ in the geometry alongside the quadratic form; rather, I really wanted finally to remove this 'inconsequence' which had always been a thorn in my side, and then noticed to my own astonishment: it looks as if this explains electricity".

Does this historical incident tell us anything about the relation of physics to mathematics? It is definitely revealing of the interaction between the two as a theory emerges and it may tell that, in this specific case, two theories seemingly irrelevant to each other, a discovery that we might think of as accidental, and the persistence of one brilliant mathematician in his wrong idea, all contributed to the instigation of what turned out to become the most fruitful physical theory of the second half of the twentieth century. But then, just one idea -even more so, when this idea is a wrong one- cannot be held responsible for any progress in physics by itself. What we shall see shortly, though, is that, at least in this specific episode in the history of physics, there has been a dynamic relation between physics and mathematics, an exchange of ideas between theoretical physicists, phenomenologists and mathematicians. It all started with the quest for the holy grail of unification -unification of the two then known fundamental forces was the leading principle for scientists in the early twentieth century. But then, there was more than just an all encompassing geometry that was required, as the Einstein-Weyl debate shows us. And the necessary re-

[18] As we have seen above, the "inconsequence" Weyl refers to is the fact that in Riemannian geometry the magnitudes of parallel transported vectors are path independent, in contrast to their directions. Apparently, Weyl considered this to be a residue of Euclidean geometry that prevented Riemannian geometry from being truly infinitesimal.

quirement was that the theory corresponded to nature. How Weyl's idea was modified and what amendments were made to it will be the topic of the following section.

1.3 The Metamorphosis of Weyl's Idea

Weyl's original idea of unifying the two forces using scale invariance was wrong, despite the fact that it was bold and appealing. It was bold because it introduced the new concept of gauge and appealing because of its unifying effect. O'Raifeartaigh, in his definitive history of the development of gauge theories, considers the choice of the word gauge to be "quite appropriate since the scale factor attached to the metric changed the measurement of length and the word gauge was in common use for measurements of length, e.g. the width of railway tracks"[19]. But then, after Weyl's idea of gauge turned out to be wrong and the concept 'metamorphosed' into something different, as we shall see, the name remained the same and some may even claim that it is misleading. In any case, this is quite common in physics, where examples of similar cases abound. Such examples are the word *mass*, whose content changed from the classical into the relativistic one in the first two decades of the twentieth century, and the word *field*, whose meaning changed dramatically in the last half of the same century. Weyl's theory was appealing despite its falsity because it did manage to unify by treating gravity and electromagnetism in the same way: both interactions arose as a result of some invariance. A very strong point in favor of his idea was the fact that through Noether's theorem, the scale invariance led to conservation of the correlate of electromagnetic current, while invariance under spacetime transformations led to

[19] O'Raifeartaigh, *The Dawning of Gauge Theory*, p.42.

conservation of energy-momentum. Although very different from the phenomenological point of view, both laws had been given a common geometrical basis and this prospect of unification appealed to Weyl. It was probably because of this appeal that the idea was not completely forgotten, despite the fact that Einstein's objection showed that the theory had no correspondence to how things appeared to be in the real world.

As years went by, quantum mechanics' advent changed things in physics' landscape and along with it the perspective of physicists also changed. The first one to relate Weyl's scalar factor to something else was Schrödinger. In a 1922 paper, Schrödinger noticed that the exponent Weyl's non-integrable factor became quantized in systems that satisfied the Bohr-Sommerfeld quantization conditions. Schrödinger then suggested that the quantization unit, which he called γ, was equal to $i\hbar$. This choice was fine in terms of units -it has dimensions of action as it should- and would restore the experimental single-valudness of the scale, the lack of which was what doomed Weyl's idea. "Strangely enough" writes O'Raifeartaigh "Schrödinger does not refer to his 1922 observation in his classic 1926 papers, but that it played a role in his invention of wave mechanics is known from a letter[20] that he wrote to London in 1927"[21].

London, who was aware of Schrödinger's 1922 paper, took Weyl's idea about the scale factor and Schrödinger's idea of its new application a step further and showed, in his 1927 paper, that in the presence of an electromagnetic field, the wave function should acquire a phase factor, which was nothing other than the transmuted Weyl factor. The general message of London's paper was clear: "the actual problem was not the presence of

[20] V. Raman and P. Forman, *Hist. Studies Phys. Sci.* **1** (1969) 291.

[21] O'Raifeartaigh, *The Dawning of Gauge Theory*, p.79.

Weyl's non-integrable scale factor but the fact that, according to Weyl is should be real and applied to the metric. If it was converted to a phase-factor and applied to the wave function instead the problem was removed. In fact, London's rather cumbersome argument was not really necessary and his proposal can be summarized by saying that in the presence of an electromagnetic field, the wave function should acquire a phase factor,

$$\psi \rightarrow e^{\frac{ie}{} A_\mu dx^\mu} \psi$$

Thus the Weyl factor, which by 1927 had been abandoned even by Weyl, acquired a new lease of life as the London phase factor"[22]. Although the factor that gave the correct theory was a phase factor, rather than the original scale one, what made the former a recognized successor of the latter was the fact that it too gave rise to coupling terms and conserved quantities as a result of applying the same variational principles and similar requirements for covariance of the resulting theories under local transformations.

Sometime between Schrödinger's 1922 idea about the scalar factor becoming a phase factor and London's 1927 idea about applying it not to the metric but to the wavefunction came Schrödinger's 1926 papers on how to introduce electromagnetism in wave mechanics. Schrödinger generalized the relativistic electromagnetic Hamilton-Jacobi equation to the relativistic electromagnetic Klein-Gordon equation by replacing the variables of the former by operators acting on the wave function. Although he did not mention it, by doing so Schrödinger was employing what is usually called the *minimal principle*[23] and though he did not emphasize the role of gauge invariance in the resulting theory, others did. These

---

[22] Ibid., p.81.

[23] The minimal principle is the principle by which the effect of an electromagnetic field on a particle of charge e is obtained by changing p_μ to $p_\mu + eA_\mu(x)$. See, for example, O'Raifeartaigh, p.17.

were first of all Kaluza (1921), who anticipated the other attempts, Klein (1926), Fock (1926, 1927) and most notably Dirac (1928).

Kaluza, Klein and Fock attempted a generalization of Einstein's theory that included a fifth coordinate. By considering five coordinates, instead of the four for spacetime, they arrived at quantum-mechanical equations for particles in electromagnetic fields that take the form of geodesic equations. But in doing so they had, of course, to reduce the five dimensions into four spatiotemporal ones and faced major difficulties in explaining why the fields should not depend on the fifth coordinate, which was transformed out of the picture. Moreover, their theory did not yield any new predictions and left the gravitational and the electromagnetic coupling constants unrelated; hence their idea remained marginal. Regardless of the original failure of the idea of dimensional reduction to blossom into a successful theory, it is worth mentioning that it did play a role in London's discovery of the successful interpretation of Weyl's original idea; after all their gauge transformations in spacetime may be considered as transformations in higher dimensional spaces, as we shall see later in this thesis. Moreover, the idea has been revived and applied in two major areas of modern theoretical physics, namely phase transitions and string theory. The success of the application is such that it makes one wonder if the history of physics is repeating itself in a way, and something similar to what happened in the case of Weyl's idea is happening here as well. In both cases the resulting theories are so successful in explaining and so far reaching in their predictions that it makes it difficult to believe that the relation between the original, mistaken ideas and their final successful reformulations is a mere accident. Rather,

it seems that in both cases the general direction was right from the beginning although the first turning taken was not.

Dirac, on the other hand, begins with the free equation for spinors with half integer spin

$$(\gamma^{\mu}\partial_{\mu} + m)\,\psi(x) = 0$$

and by using the minimal principle, namely by substituting $\partial_{\mu} \rightarrow D_{\mu} = \partial_{\mu} - ieA_{\mu}$ he derives electromagnetic interaction terms.

Weyl himself took the new applications and interpretations of his old idea a step further, in his 1929 paper, and developed a complete theory out of it. A striking similarity between the theories of electromagnetism and gravitation is that charge conservation in the first and energy-momentum conservation in the second are derived in the same way: by requiring invariance of the theory under certain variations. This similarity was enough to convince Weyl that the two are closely related and to drive him to the complete and explicit formulation of the analogies between the two theories by means of the tetrad formalism in 1929. Moreover, by adopting London's reinterpretation of the non-integrable scale factor of the metric as a non-integrable phase factor of the wavefunction, he was able to overcome the objection that threatened to abolish the original theory, but he also went a step further by proposing that electromagnetism is derived from the gauge principle. This idea proved to be extremely fruitful later on, in the study of weak and strong interactions. Here is a summary of what Weyl did in that paper.

First of all he introduced the concept of the two component spinor in a different way than that of Dirac. In this mathematical framework he discussed time reversal -his spinors'

theory violated time reversal- and violated parity as well and though at that time parity violation was out of the question, later on it turned out to be true. In order to integrate the two component spinor theory with gravitational theory, Weyl followed Wigner's idea of using local tetrads, a concept that had been introduced by Einstein not long before. The tetrad formalism is very useful because it does not only allow for handling spinors on curved spaces but also it allows for deriving the energy-momentum conservation laws and it makes the analogy between electromagnetism and gravity manifest. Moreover, given that each tetrad has sixteen degrees of freedom -count ten for the Riemannian metric and six for the Lorentz group- the tetrads are determined by the metrics up to a local Lorentz transformation. This formulation allows for an algebraic treatment of differential geometry and a major advantage is that it exhibits the resemblances between gravity and gauge theories. Then, Weyl discusses spinors in curved space and although he did not mention Noether and her theorems in his paper, he applied them in his tetrad formalism to derive the conservation laws for their momentum, both linear and angular that result from invariance under coordinate transformations and internal Lorenz transformations of the tetrad respectively. Then he expressed gravity in the tetrad formalism so that the analogy between electromagnetism and gravity became apparent. Finally, he went on with the derivation of electromagnetism from what is now known as the gauge principle.

In this last part of his paper Weyl takes three steps. The first one justifies the rigid (global) phase invariance of the spinor theory on the basis that the spinors are defined as representations of the $SL(2,C)$[24] which is a subgroup of the $GL(2,C)$, hence the intrinsic

[24] 'C' stands for complex.

gauge freedom in the spinor theory which does not distinguish between $\psi(x)$ and $e^{ia}\psi(x)$. The second step explicates that as it is 'natural' to generalize from the rigid Minkowski tetrad to a local tetrad, so it is to generalize from a rigid a to a local $a(x)$ that allows for $\psi(x) \rightarrow e^{ia(x)}\psi(x)$. The exponent here is independent of the tetrads and this manifests the fact that the locality of the phase parameter is intrinsic. In the third step, the gauge principle is used to obtain electrodynamics. According to it, it is required that a theory with an action invariant under a rigid phase transformation remains invariant when the transformation becomes local in way similar to that of diffeomorphisms. Namely, just as when requiring invariance under local diffeomorphisms the derivative ∂_μ should change into the covariant derivative $\Delta_\mu = \partial_\mu + \Gamma_\mu(x)$, so when requiring invariance under the local $U(1)$ group the derivative Δ_μ should be modified accordingly: $\Delta_\mu \rightarrow D_\mu = \Delta_\mu - \frac{ie}{c}A_\mu(x)$, where $A_\mu(x)$ is the connection of the Abelian group, also known as the gauge group. Hence, considering the gauge principle, electromagnetic interactions are derived from a geometrical principle, just like gravitational interactions[25]. What was particularly appealing to Weyl was the fact that this time round the principle of gauge invariance "derives not from speculation but from experiment"[26], whence his new brain child was no longer vulnerable to the criticism that it does not agree with nature.

So what does the term gauge mean, after all, and what are its appropriate uses? As we saw, in the 1918 Weyl paper, where the term was first introduced, it had a meaning and an application very similar to its every day use; it was a (symmetry) scale factor of

[25] Notice, however, that the problem with the gauge principle is that it is only an assumption, because although if a theory is invariant under local gauge transformations is also invariant under global gauge transformations, the inverse is not necessarily true. Later on in this thesis we will get back to this point.

[26] Weyl, 1929.

the metric and hence it affected the scale of length measurements. But since then, the scale factor metamorphosed to a phase factor and thus the meaning of the term lost its relevance. However, the term itself survived in the notions of gauge symmetries, gauge transformation and the gauge field or simply the gauge. When we talk about gauge symmetries in the context of theoretical physics we mean symmetry transformations that leave the action of matter and interactive fields invariant; these may be related to either spatiotemporal transformations or transformations of internal degrees of freedom and although only in the latter there is a phase factor involved, they all give rise to interaction terms by using the so called covariant derivative and the expected conservation laws as a result of Noether's theorems. The fact that spatiotemporal diffeomorphisms do not make any use of phase factors multiplying the wavefunction makes them look different from the other gauge transformation that definitely deserve the name gauge and it poses questions about how legitimate it is for these transformations to be considered as part of the gauge family.

One point we want to clarify here is that although the presence of initially a scalar and later a phase factor worked as a heuristic assumption at the beginning of the gauge theories, the truly crucial elements that probably have been guiding principles for Weyl and the others were the similarities in the description of gravity and electromagnetism, namely the derivation of conservation laws and the restoration of invariance -and manifestation of coupling terms that could be interpreted as interactions- when the 'flat' derivative was replaced by the covariant derivative. If we define gauge symmetry to be a symmetry that involves a phase factor multiplying the wavefunction, then gravity and diffeomorphism invariance have no place there. But to our view, which is also the view of physicists and

mathematicians that have worked on these fields since their discovery, the most important aspect common in both is the presence of an arbitrary function, the connection; hence the term has been and should be broader than that. Therefore, gravity can be considered as a gauge theory provided we bear in mind the broader picture. In the chapters to follow, we will take a closer look at gauge theories, in order to clarify and expand on what we mean by the term. But in the last part of this chapter, we would like to continue the brief historical introduction by giving a brief account of what happened after 1929, since it was in the second half of the twentieth century that the tremendous phenomenological and experimental success of gauge theories became apparent. A last remark concerning the Weyl-Einstein debate we are obliged to make here and postpone any further discussion until we get to chapter four. Einstein's objection transformed to a very successful prediction after the modification of Weyl's original idea. In view of London's reinterpretation, and fifty three years later, C. N. Yang pointed out that in the new interpretation where Einstein was talking about scales, we would now have to consider phases and while the original objection was that two rods taken along different paths would not have different scales, two electrons -the microscopic equivalent of charged rods!- taken along different paths would have different phases. This, of course, was the question that Aharonov and Bohm asked in 1959, and apparently they asked it independently and without reference to the original objection of Einstein. As we shall see, the experiments that were conducted concluded that the prediction was indeed correct and its success is highly regarded as an endorsement about the validity of the theory that predicted it.

1.4 Swimming Against the Phenomenological Tide[27]

In this section, we will attempt a brief account of what happened from 1929 until the 1980s in the world of theoretical physics. There is a very good reason, though, why we have to include it here -no matter how incomplete and sketchy. So far we have discussed the onset only of gauge theories and we have said nothing about the weak and the strong interactions that were integrated into the picture of gauge theories later. The inclusion of these two interactions in the gauge theories' picture was **the** major success of the theory and turned it into the most influential theory of, at least, the twentieth century; influential in the sense that it changed dramatically the way we perceive the world. If we do not mention the intellectual achievements of that period and the interactions, influences and dialectics in the scientific community that led to them, we will fail to get a comprehensive impression of the dynamics in the relation between physics and mathematics; the relation that led to beautiful mathematical structures that are very successful in describing the world.

Between 1929 and 1936 there was nothing new from the experimental physics front, which meant that there was no indication that the nuclear force fields might exhibit some sort of vector character and hence might be described by using gauge potentials as well. However, in 1936 Yukawa suggested that as atomic forces are mediated by photons, so the strong nuclear forces might be mediated by massive mesons. Although we do not know what gave Yukawa this idea, we think it is plausible to speculate that the existing theory and its success played a heuristic role in this case. It was possibly an argument by analogy

[27] The title of this section is borrowed from a phrase that can be found in O'Raifeartaigh's *The Dawning of Gauge Theory*, p.7. O'Raifeartaigh's book is highly recommended as a wonderful resource for more precise and complete historical detail. For a standard physics introduction to this material reference may be made to Aitchison & Hey's *Gauge Theories in Particle Physics*.

and it was a valid and legitimate one since it contained what Hesse calls neutral analogies[29] that could be tested against experiment. Those tests would decide whether the idea was correct or not.

Two years later, in 1938, Klein was pursuing further his 1926 ideas and in 1939 he presented the first attempt to generalize gauge theories so that they incorporated the Yukawa meson. Klein ended up with what we would recognize nowadays as a $SU(2)$ gauge structure and, as though this was not enough, responding to a comment by the audience he anticipated the gauge group used in the standard model by generalizing the $SU(2)$ Lie algebra of the meson fields to $SU(2) \times U(1)$. But Klein's work was forgotten and O'Raifeartaigh speculates that this happened because the paper was never published, it was only presented in the 1939 Conference on New Theories in Physics in Poland, its ideas were not appreciated by the eminent physicists that were present and the second world war occurred shortly after the paper was presented. As we mentioned before, Klein attempted this generalization by introducing a fifth coordinate, which then he had to 'reduce' and at that time the physical significance of that was not clear. So despite the fact that the dimensional reduction provided some means for constructing what later on was recognized as non-Abelian field strengths, at that time this point was not fully appreciated.

Nevertheless, ten years down the line, three independent attempts to include non-Abelian Lie groups in gauge theories appeared and apparently each one was motivated in different ways. The one that came first was that by Yang and Mills and as a matter of fact the non-Abelian gauge theories are called Yang-Mills after them. Yang, who was working

---

[29] Analogies will be discussed in some detail in chapter four.

as a graduate student on field theories, studied Pauli's review articles[31] on the subject and impressed by the two main ideas of the theory, namely that conservation of charge followed from the gauge (phase) invariance of the theory and that interaction terms arise when applying the gauge principle, he tried to generalize it to include isospin interactions. Along with Mills, they successfully constructed a non-Abelian gauge theory in 1953, which was published in their 1954 paper. As it turned out later, when the axial-vector character of weak interactions manifested itself to phenomenologists -that happened in 1958- it started to become clear that the Yang-Mills field was not appropriate for the description of weak isospin interactions but of weak interactions instead and the theory fully blossomed only when they sorted out the problem of giving mass to the connections -or gauge fields- by symmetry breaking and when it was shown that the theory as a whole was renormalizable; but these two last issues are another story. So once again here, as in the case of Weyl, agreement with the experimental results was the crucial arbiter and in the light of disagreement they had to reconsider the applicability and the application of the theory and shift it from weak isospin to weak interactions. In this case though they did not have to revise the theory.

Shaw's successful attempt that led to the same conclusions as that of Yang and Mills was inspired by a manuscript of Schwinger's. Shaw wrote about this: "[Schwinger] introduced electromagnetic interaction in this way -he used real spinors and so had $SO(2)$, rather than $U(1)$, invariance and the generalization to $SU(2)$ invariance seemed to shout it-

[31] It is worth noting that Pauli was initially one of the opponents of Weyl's idea, but finally he was enthralled by it and became one of its foremost proponents, as can be seen in his (1941) as well as in his later works on dimensional reduction (1953).

self out!"[33] He too, like Yang and Mills, was concerned with isotopic spin and noticed that the rigid $SU(2)$ invariance of it would give connection terms and a covariant derivative if it was made local. Hence, given that his PhD thesis, where his approach was first published, was dated 1955, he arrived at the same result only a year later.

So did Utiyama, who reached the same conclusions too about non-Abelian gauge theories by extending Weyl's gauge principle to general Lie groups. Utiyama's approach was more comprehensive since it included gravity, however his paper appeared later than Yang and Mill's, as well as of Shaw's. Even before 1954, Utiyama was working on general gauge theory stimulated by Yukawa's theory. Though he did give a talk in Kyoto university in 1954, he was not happy with his results because they did not seem to agree with Yukawa's -the problem of the mass-less-ness of the gauge fields, that is- and because in this case things seemed to go the other way round: in this case there seemed to be a physical law following from gauge invariance and not the other way round. So, Utiyama did not publish his paper, to his regret apparently.

In any case, by 1955 the physics community had the general formulation of gauge theories that included non-Abelian Lie groups, while at the same time the mathematics community was developing the fibre bundles formalism, a formalism that encompassed gauge theories. The development in mathematics was motivated for different reasons, namely mathematicians were interested in the study of manifolds with topological anomalies. But it took twenty more years of developments in physics -experimental and phenomenological at first and with further modifications, adaptations and alterations of the

---

[33] Shaw in a letter to Kemmer, 26th May 1982.

mathematical parts then- before the whole picture was complete. And apparently, it was only in 1958 that phenomenological evidence of the axial-vector character of the weak interactions made the dialectics between the already existing Yang-Mills theory and experiment possible. It is in this sense that those who constructed gauge theory "were swimming against the phenomenological tide"[35], and yet, they proceeded regardless! But then, this exhibits one of the biggest strengths of a successful theory: it probes and anticipates and predicts and guides.

Why does this happen? How does this happen? We do not propose a complete answer in the present work. But one thing is certain, along with physical and mathematical intuitions, that we cannot explain how exactly they arise, experiment and agreement with it lie at the heart of this amazing structure that is called gauge theories. It was precisely this requirement of agreement with experience that inspired Einstein's justified criticism of Weyl's original idea and it was thanks to this criticism that Weyl realized he had taken a wrong turning. On the basis of this criticism the idea was then successfully transformed.

1.5 A very Brief History of Fibre Bundles

The results of Weyl's ideas were far reaching in physics, as we have seen, but what is even more amazing is that they did not influence physics only; they also motivated progress in an area of mathematics currently known as *fibre bundles*. Fibre bundles is a branch of differential geometry, it is the mathematical tool that is extensively used in the description of gauge theories in physics and, roughly speaking, deals with manifolds and symmetry

[35] O'Raifeartaigh, p.7.

groups acting on those manifolds. It is widely considered to be the most appropriate mathematical formalism for the description of elementary particles (or shall we say fields?) and fundamental forces, whether they are described using non-Abelian gauge symmetries or other more elaborate physical theories, like for example string theories. The fibre bundle formulation of gauge theories is a fairly recent development in theoretical physics, it only dates back to the mid 1970's. As a matter of fact, it seemed as though physicists 'came across' a ready made formalism, that of fibre bundles, after they had discovered and developed gauge theories independently. Quite miraculously, it seemed, they realized that a formalism that suited their needs and purposes was already there and so they adopted it. But the truth, although hidden by the debris of the several incidents that mark scientific discovery, is that non-Riemannian geometries, in general, and fibre bundles, in particular, were inspired by physics and developed in parallel with, albeit faster than, gauge theories. Although the routes of the two enterprises were not always connected, and at times were even independent, they crossed again and again. In this section we will briefly delineate the route that led to the fibre bundles and to their deployment by the theoretical physicists and we will try to reveal the interactive relation between the two, the physical gauge theories and the mathematical fibre bundles.

As we have seen, Einstein's theory of general relativity inspired Weyl to produce a geometry that would accommodate both gravitational and electromagnetic interactions in a unified way. Although Einstein's theory was based on Riemannian geometry, it nevertheless inspired Levi-Civita, at first, and Weyl and Cartan shortly afterwards, to pursue the notion of parallel transport further, so that it did not contain "a residual element of rigid ge-

ometry"[37]. In 1917, one year after Einstein's theory of gravitation, the mathematician Levi-Civita introduced the concept of parallel transfer. Inspired by "la grandiosa concezione di Einstein", apparently, Levi-Civita realized that the covariance of the Riemannian derivative and the Riemannian tensor was not due to the fact that the Christoffel connection was derived from the metric; rather, the covariance was the outcome of the transformation properties of the Christoffel connection with respect to coordinate transformations. That fact declared the status of the connections as independent entities. Weyl was the one who introduced the notions of connection and parallel transport to physicists through his 1918 paper and his later works, while Cartan was one of the mathematician-pioneers of what became modern differential geometry. Very enthusiastically, O'Raifeartaigh points out that "the significance of the Levi-Civita-Weyl-Cartan development can hardly be overestimated. From the point of view of mathematics, it liberated Riemannian geometry from the metric and thus opened the way to a much more general concept of differential geometry, with the emphasis on differentiable manifolds and on their topological properties. This led to a sustained mathematical development which culminated about 1950 in the theory of fibre bundles. [...] From the point of view of physics, the Levi-Civita-Weyl-Cartan development paved the way for a geometrical understanding of electromagnetism and the weak and the strong interactions and for understanding their common structure"[38]. The connection in gravity is related to the derivative of the metric, but in the rest of the fundamental interactions it is defined independently and it represents the interacting field, as we shall see when we present the formalism. However, thanks to Weyl's idea and to its subsequent extensions

[37] Weyl, 1918.

[38] Ibid., p.40.

by Utiyama, Yang and Mills, it became known that even in the case of the other interactions, the covariance is the outcome of the transformation properties of the gauge field with respect to local phase transformations and the gauge field itself is related to parallel transport. Hence the mathematical research on the relations between the properties of a space and the properties of the symmetry groups acting on that space was bound to be relevant to these theories too.

1.5.1 From *Sphere Spaces* to *Sphere Bundles* to *Fibre Bundles*

In mathematics, then, during the 1920's and the 1930's there was work going on in the areas of symmetry groups, topology and differentiable manifolds. Along with Elie Cartan, who by 1929 had become aware of and appreciated the fact that -what he called- the invariant integrals of certain homogeneous spaces were related to topological properties of those spaces[39], C. Ehresmann, H. Hopf and H. Whitney were also becoming aware that "[t]he properties of a homogeneous space in which acts a Lie group simply expresses the properties of this group"[40]. As Ehresmann points out in his 1934 paper, "[i]t would be very interesting if we knew the relations between the topology of such a space and the properties of its structure group"[41] but their knowledge on the subject at the time was limited. "In the mean time, in his research concerning simple groups and homogeneous symmetric spaces, Mr. E. Cartan has reached remarkable conclusions/results that reveal some of

---

[39] See *Sur les invariants intégraux de certain espaces homogènes clos et les propriétés topologique des ces espaces* (Ann. Soc. pol. Math., 8 (1929) 181-225).

[40] C. Ehresmann, *Sur la topologie de certain espaces homogènes*, Annals of Mathematics, 35, no.2, 1934, 396-443.

[41] Ibid.

these relations"[42] and with these conclusions he was laying the foundations of what was to become modern differential geometry and the fibre bundles approach. Ehresmann's paper continued to investigate the topological properties of such spaces, while on the other side of the Atlantic, Whitney was publishing a year later (1935) a paper where a direct ancestor of the fibre bundles first appeared, under the name *sphere spaces*. In the opening section of this paper, Whitney wrote: "Spaces often occur in which points themselves are spaces of some simple sort, for instance spheres of a given dimension. The set of all great circles on a sphere is such a space. Some general types of sphere-spaces are given in §3 below, and some specific examples in §8. Locally, sphere-spaces are product spaces; but in the large, this may no longer hold. In this note we define invariants which serve to distinguish different sphere-spaces when they have the same 'base space' "[43]. One of the examples of sphere-spaces he gives is what we now recognize as the tangent bundle.

By 1939, and while mathematicians like Hopf were investigating the relations between the topology and differential geometry of analytic Riemannian manifolds, Feldbau published the paper *Sur la Classification des Espaces Fibrés*, where, for the first time, appears the term *fibrés* (adjective), which will be adopted as *fibre* (noun) in English. Ehresmann and Feldbau, in a joint paper that was published only two years later, give the first definition of a *bundle*, a definition that had not yet discarded references to coordinate functions and equivalence classes. By 1940, Whitney had renamed his sphere-spaces as sphere-bundles and he also defined the term fibre bundle.

---

42 Ibid.

43 H. Whitney, *Sphere-spaces*, Proc. Nat. Ac. Sci., 21 (1935) 464-468.

The first monograph on fibre bundles came under the title *The Topology of Fibre Bundles*, was written by Norman Steenrod and it was published in 1951. In his introduction, Steenrod calls attention to the fact that "[t]he literature is in a state of partial confusion, due mainly to the experimentation with a variety of definitions of 'fibre bundle'. It has not been clear that any one definition would suffice for all results"[44]. What Steenrod attempted to do with his monograph was to provide an organization of the material that had been published between the years 1935-1948, and gave in it the first direct definition of a fibre bundle that avoids coordinate functions and equivalence classes.

Apart from the work on fibre bundles, at the same time there was research done on Lie groups, differential forms and connections, an area that naturally became part of modern differential geometry. We will not attempt even a brief historical account of the subject here, but we would like just to mention an acknowledgment to the contribution of Weyl in the subject. Claude Chevalley dedicated his 1949 book *Theory of Lie Groups* to Elie Cartan and Hermann Weyl. Although there are hardly any references in the text, in the introduction he clearly states that certain of the ideas of the book have been inspired by the two mathematicians. We are mentioning this at this point as a reminder that, basically, this very powerful mathematical theory, modern differential geometry, has got certain of its fundamental ideas traced back to mathematicians like Levi-Civita and Weyl who were interested not just in advancing a mathematical theory per se, but in developing a theory that might find applications in physics. There was some strong physical intuition hidden behind the work of these mathematicians.

---

[44] N. Steenrod, *The Topology of Fibre Bundles*, v.

By the end of 1960s the mathematical theory had been completed and the two volumes of Kobayashi and Nomizu, *Foundations of Differential Geometry*, resolved any possible disputes about definitions and terminology. Nevertheless, the physics' front was advancing at a slower pace and hence the fact that the two were following parallel routes did not become apparent but only since the mid 1970s. The main reason why physics was lagging behind, at least so far as the fundamental forces except gravity were concerned, was lack of experimental input. Let us not forget that for years physicists did swim against the phenomenological tide. But when the time came, differential geometry and fibre bundles were used and proved to be very successful and heuristically extremely fertile, to the extent that they amazed everyone in the scientific community as well as in the mathematical and the philosophical ones. And the question that arose then was: what's the relation between the two? This question we will try to answer from a philosophical perspective in the following chapters, but the main thing to remember from this brief historical introduction is that both gauge theories and differential geometry share some ideas as part of their origins.

1.6 The Aftermath

As we have already seen, gauge theories cropped up from an idea, an intuition, that occurred in the mind of a mathematician and originally it was wrong: it did not agree with experiment. Since the pursuit was not purely mathematical but in relation to physical problems, the theory could have been discarded. However, the theory was hammered but not destroyed by experimental considerations and the dynamic interactions between its authors

and others, that took place in a period of eleven years, put the heart of the idea into the appropriate mathematical framework and shaped up an attractive theory[45].

Something similar happened in the second part of the development of gauge theories; similar with respect to the interactions between mathematicians, theoretical physicists and phenomenologists. Similar, yet not the same because no two incidents in the history of physics and mathematics are exactly the same. In this case too, at the beginning the physicists produced an extension to gauge theories that did not agree with experimental data and phenomenological propositions. Nevertheless, nature rather than humans, this time, provided further evidence that theorists were on the right tracks. Further experimental evidence and further theoretical adaptation -but not metamorphosis- were required before Glashow, Iliopoulos and Maiani, at first, and Weinberg and Salam, finally, developed the standard model for the electroweak interactions. And then more interactive work between experimental and theoretical physicists and mathematicians took us to unification theories that included the strong interactions and the fibre bundle formalism. And the story continues with greater unification schemes that aim to include gravity in the picture in a non-problematic way.

The similarities in the two phases of the creation of gauge theories are the following. In both cases it all started in disagreement with the known phenomenology. Intellectual interaction clarified the situation and put things right, so that the final theoretical result was in agreement with observation. The difference is that while in the first phase the initially

[45] We mention in passing here that in at least one more case this kind of interaction between a mathematician -Emmy Noether- and physicists led to an extraordinary and very influential (in physics) piece of mathematical work, namely Noether's theorems, at around the same period. Although her name is not particularly referred to in the works of Weyl and others who worked on field theories, it is almost certain that there was communication and interaction between them.

proposed theory had to undergo a partial, though substantial, transformation in order to match, in the second phase the theoretical basis was already established and what was important was the input of new data mainly from the experimentalists' side. Yet in both phases -and probably in all successful theories in physics- the mathematical ideas are based on a physical consideration -there was physics in the heart of the idea of the connection.

Chapter 2
Mathematical Representations of Physics

The relation between science and mathematics has always been a very successful and fruitful one, yet at the same time, one that raises several philosophical questions, mostly with non-conclusive answers. The success of this long term relation makes it almost necessary for one to admit that the fact that mathematics describes, explains and even predicts, physical matters of fact is not an accident. However, at least so far as physics is concerned[46], mathematics by itself cannot give an adequate explanation of a physical event, for that reason some linkage is needed. Roughly speaking -and for the time being let it be like that- the prevailing suggestion among the physicists is that this linkage, this connection is provided by the interpretations of the theory. But then, the question that arises is, what do we mean by interpretation? In this chapter we are investigating precisely this relation between mathematics and physics, and one thing we are arguing for is that so far as theoretical physics is concerned, *the nature of this relation is one that does not allow us to separate completely the mathematical from the physical aspects of a theory.* The two are so inextricably entangled that one cannot strip a physical theory of its mathematics and just keep the physics because as we shall see, one does not know exactly where to draw the line between the two. In chapter 4 of this thesis, we will return to the issue of interpretations and examine it within the context of gauge field theories.

---

[46] Here we are only concerned with the relation between physics and mathematics and not the rest of science.

2.1 The Mathematical and the Physical

2.1.1 Raising the Issues

In his book *Thinking about Mathematics*, Shapiro distinguishes two major questions that those concerned with the relation between physics and mathematics should tackle. The first one is a 'how' while the second is a 'what' question.

"How is mathematics applied in scientific explanations and descriptions?"[47] is the first question and Shapiro, to clarify things, talks about applications of two different types of mathematical entities, namely, *concepts* and *theorems*. Since "we apply the concepts of mathematics -e.g. numbers, functions, derivatives, integrals, Hilbert spaces etc.- in describing non-mathematical phenomena"[48] and "we apply the theorems of mathematics in determining facts about the world and how it works"[49], our how-question could fork into two. How are mathematical concepts applied in scientific explanations and descriptions of non-mathematical phenomena? How are theorems applied in deducing and/or determining facts about the world and how it works?

The what-question is phrased by Shapiro as follows: *"what is the philosophical explanation for the applicability of mathematics to science?"*[50] Or, in other words, what is the philosophical explanation for the applicability of mathematical concepts in explanations and for the applicability of theorems in deductions (which could be perceived by many

---

47 Shapiro, 2000, p.36.

48 Ibid.

49 Ibid.

50 Ibid.

as explanations too). In this thesis we will not deal with the what-question, but we will attempt to give some answers to the how-questions. These answers will be based mainly on conclusions drawn from the application of differential geometry in field theories.

A somewhat different, yet compatible with Shapiro's, classification of the problems related to the application of mathematics is based on Steiner (1995), who recognizes problems of meaning (or semantics), problems about the relation of mathematical to physical objects (or metaphysical) and problems about physical reality and mathematical objects (or how physics relates to mathematics).

The first type of problems is about interpreting mathematical terms. In scientific explanations, especially in physics, we use both mathematical and physical terms. The mathematical terms employed ought to be interpreted in such a way that they have some sort of physical meaning per se or as they are used in mathematical proofs and derivations, so that they become relevant and meaningful in scientific descriptions, explanations and predictions. Once we have interpreted the mathematical terms, we are able then to use them directly in derivations that we label scientific, rather than mathematical. We will get back to this issue in chapter 4, where we will explore possible interpretations of the fibre bundle formalism, a purely mathematical 'construction' which is used in gauge field theories.

The second type of question arises if we presuppose that there are mathematical as well as physical objects and that they are distinct. Then the challenge we face is to account for the nature of mathematical objects that allows them to relate to and apply in the physical world.

Finally, if we reverse subject and object in the last type of question, we express the last sort of problem related to the applicability of mathematics to science. Namely, the issue now is to account for the nature/properties (or what else do we call them?) of the physical world that makes specific concepts and formalisms of mathematics so applicable to it. Some more specific questions that could be asked within this context, as Shapiro put it, are these. "What is it about the physical world that makes arithmetic so applicable? What is it about the physical world that makes group theory and Hilbert spaces so central to describing it?"[51] According to Steiner, for each concept and every successfully applied formalism we should expect a different answer.

Without too much emphasis on the second type of problems and with Steiner's last remark in mind, in this chapter we will attempt to shed light on certain properties of some physical objects that allow for specific applications of mathematical concepts and formalisms. From the various philosophical approaches to the relation between mathematics and physics, we will focus on two: Field's programme and M. Redhead's structuralist ideas of *surplus structure.* The reason we chose these two approaches is that they both deal plainly with representation of physics by mathematics. Field's main thesis is that at least in principle, it is possible to reformulate physical theories so that the mathematical entities are avoided, while Redhead's is that not only this is not possible, but also it is the purely mathematical surplus structure that 'controls' the physical, as we shall see. Hence we will discuss Field's programme and criticisms against it, and then M. Redhead's. From this per-

---

[51] Ibid.

spective we will then investigate how the notion of *symmetry* is applied and what we can get out of this application.

Each of these problems may occur on several levels. So, we may ask how it is that a particular mathematical fact can serve as an explanation of a non-mathematical fact, or what is the relevance of a given mathematical/scientific theory as a whole, or why is the entirety of mathematics essential to science. In this chapter we are discussing issues related mostly to the second level -we will only touch upon the third- while in chapters 3 and 4 we will also focus on the first, discussing particular facts.

One last remark before we move on. Shapiro points out that "occasionally, areas of pure mathematics, such as abstract algebra and analysis, find unexpected applications long after their mathematical maturity. Mathematicians have an uncanny ability to come up with structures, concepts and disciplines that find unexpected application in science"[52]. This is yet another issue that can be illustrated by several examples from the history of science. A notable example that is related to this thesis is the development of the fibre bundle formalism of differential geometry that found applications in physics almost two decades after it reached its maturity in the minds and the interests of the mathematicians. Once again, we will come back to this point in the last chapter of this thesis, after we have developed the fibre bundle formalism, anticipating to get some insights on how the relation between physics and mathematics developed, at least in that specific example.

---

[52] Ibid., p.39.

2.1.2 The Question of Choice: Which Mathematical Representation and Why?

As is well known, a physical theory may have more than one mathematical representation. This problem we call *ambiguity of representation* and examples from physics abound. Take for instance the case of classical mechanics. There we are accustomed to using Euclidean geometry, but other metric geometries, like for example Riemannian geometry, would do as well. The question, hence, is which one to choose and on what grounds. Nagel, in *The Structure of Science,* puts forward two different attitudes towards answering this question. The one he supports is known as *conventionalism.* Conventionalism advocates that if Euclidean and Riemannian geometries are like languages which are intertranslatable into each other, "the sole difference between the two systems of statements obtained in this way is that the *same facts* receive *different formulations*"[53]. So, "as far as the empirical facts to be codified and predicted are concerned it will make not an iota of difference which language to adopt. However, we may find one language more convenient than another, perhaps for several reasons"[54]. On the other hand, if we consider that the two different systems of statements are mutually incompatible, "the above question can now be taken to mean 'Since the alternative applied geometries cannot all be true, is there any way of deciding between them, and are there any considerations based on the empirical facts that make the adoption of one system quite compelling?' "[55]. To answer the question in this case, one has to identify the geometry that is true and this should be based on empirical facts only. Yet such an

---

53 E. Nagel, *The Structure of Science*, p.253.

54 Ibid., pp253-4.

55 Ibid., p.254.

inductive step gives rise to several problems that plague this one, along with any other realist approach. For the purpose of this thesis, we will not expand on these two approaches. Nevertheless, we will see what could be said about the issue of ambiguity of representation in Field's nominalist programme, in Shapiro's structuralist approach and in Redhead's 'second order' structuralism.

Aside from this type of ambiguity in the representation of physical theories, we also encounter another one, *the ambiguity that a specified representation allows for, within the* same *mathematical representation.* This second type of ambiguity is one that has physical import and we will discuss it in more detail later on in this chapter. But for now, we will go back to the question *"How is mathematics applied in physical explanations and descriptions?"* and discuss some of the attempts to answer it.

2.2 Field's Idea

Field's idea, by and large, was that in doing science we can dispense with numbers, which are nothing more than a conservative extension of the theory itself. This view he calls *nominalism.* His programme for nominalizing science has been criticized and Shapiro has shown it to suffer the same faux pas as Hilbert's programme for mathematics, to which it is structurally analogous. For Hilbert, the basis was finitary mathematics, the instrument was ideal mathematics and the necessary condition was consistency. In Field's programme, on the other hand, the basis is nominalistic science, the instrument used is mathematics and the necessary condition is conservativeness. Hilbert's programme suffered a severe blow from Gödel (1931, 1934) and his incompleteness theorem, while Field's attempt was found

to be non-conservative when Shapiro (1983) discovered a counterexample, a sentence G in the nominalistic language that could be derived within the extension at the same time that it was not a theorem of the synthetic physics.

Despite the problems, Field's book is admittedly "one of the few serious, sustained attempts to show how mathematics is applied to sciences"[56]. For this reason, we are presenting, examining and adding to the criticisms against his idea.

2.2.1 Science Without Numbers: a Defence of Nominalism

In his introduction, Field defines nominalism as"the doctrine that there are no abstract entities"[57], like for example numbers, functions, sets, or any similar entities. Since such entities do not exist, the argument goes, it is not legitimate to use such "terms that purport to refer to such entities, or variables that purport to range over such entities, in our ultimate account of what the world is really like"[58]. Taking into consideration another assumption, namely that physical theories describe the world the way it really is, we then face a problem. The problem is that, as a matter of fact, in developing physical theories one has to use mathematics and along with mathematics, references to and quantifications over the kinds of objects that are not supposed to exist. So, how is it possible that we give an ultimate account of what the world is really like if we use in our account entities that do not really exist?

---

56 Shapiro, *Thinking about Mathematics*, p.237.

57 Hartry H. Field, *Science without Numbers*, Princeton University Press, 1980, p.1.

58 Ibid.

A popular resolution among the nominalists is to actually interpret the mathematics involved in physical theories so that the mathematical terms involved do not make reference to 'forbidden', abstract entities, but only to other types of entities, like for example physical objects, linguistic expressions and mental constructions. Field's approach, however, is different. As he writes, "I do not propose to reinterpret any part of classical mathematics; instead, I propose to show that the mathematics needed for the application to the physical world does not include anything which even *prima facie* contains references to (or quantifications over) abstract entities -and this includes virtually all of conventional mathematics- I adopt a fictionalist attitude: that is, I see no reason to regard this part of mathematics as *true*"[59].

To do so, Field introduces the notion of conservative extension, and while he outlines his strategy, at the same time he tries to counteract the already existing arguments against the nominalist position. The task that the advocates of his position face then is that of reformulating all of science, in general, and physics, in particular, so that it does not refer to nor does it quantify over abstract entities. Field's attempt has been criticized in particular by Shapiro who in his 1983 showed that the idea of mathematics as a conservative extension of a theory fails. But before we proceed to the criticisms of Field's programme, let us outline the programme itself.

It is worth noting that the book is a long reductio ad absurdum against the Quine-Putnam indispensability argument. The indispensability argument, roughly, states that since mathematics is essential for science, it must be true and since it is true we should

---

[59] Ibid., p.2.

believe in the existence even of the abstract entities that it involves.[60] Hence, Field begins with the assumption that standard mathematics is correct and attempts to show that, nevertheless, mathematics is not indispensable to science.

On the other hand, the main argument for nominalism is the so-called epistemological argument and in Field's formulation it runs as follows. What we may call 'the reliability thesis' claims that when mathematicians believe a claim about mathematical objects, then the claim is true. If the reliability thesis is true then it must be explained. But the reliability thesis cannot be explained, therefore is not true. This 'destructive' argument only manages to justify -not without controversy, of course- why the nominalists would not want to retain current theories. However, it does not provide any motivation for embarking on a project of reconstructing mathematics in a nominalistic way. This motivation is to be found as a response to the above mentioned indispensability argument, which, according to Burgess and Rosen, implies that we should believe in abstract entities only because we do not have nominalistic alternatives to current scientific theories and hence "it makes a major concession to nominalism, essentially the concession that if nominalistic alternatives to standard scientific theories could be developed, then they should be adopted"[61].

In the first chapter of his book, Field tries to establish that while mathematics does not yield genuinely new conclusions about observable entities, physical theories do yield genuinely new claims about observables. To do so, physical theories make use of theoretical entities, however, these theoretical entities are dispensable, he argues. His first task

---

60 For a detailed discusion of the argument see, for example, *Putnam's Philosophy of Logic.*

61 Burgess & Rosen, *A Subject with no Object, Clarendon Press*, 1997, p.64.

is, therefore, to demonstrate that the utility of mathematical entities is different from the utility of physical entities, and here is how he does it.

2.2.2 In What Ways 'Utility of Mathematical Entities' is Different from 'Utility of Theoretical Entities'

Field argues that if logic does not yield genuinely new conclusions, we can give a clear and precise sense to the idea that along the same lines "the part of mathematics that does make reference to mathematical entities can be applied but without yielding any genuinely new conclusions about non-mathematical entities"[62]. For him, the only reason why mathematics is important relies on the fact that it is truth preserving and therefore it can be used to deduce consequences from premises. However, mathematical entities are dispensable in the following sense. Consider that a nominalistic assertion is one that makes no reference at all to abstract mathematical entities, then "if you take any body of nominalistically stated assertions N and supplement it with a mathematical theory S, you don't get any nominalistically-stateable conclusions that you wouldn't get from N alone"[63]. On the other hand, theoretical entities that appear in physical theories play an essential role in them and in the deduction of a wide range of phenomena from them, he claims, and since there are no alternative theories known that make no use of similar entities, they are indispensable to them.

In order to show that mathematics is conservative, Field points out that number theories or pure set theories are of no interest, since they do not apply directly to the physical

---

62 Shapiro, *Thinking about Mathematics*, p.16.

63 Field, *Science without Numbers*, p.9.

world, in other words they do not enable us to deduce nominalistically-stateable consequences from nominalistically-stateable premises. However, in order to make use of this attribute of mathematics, he requires some sort of bridge between the pure objects of the world and the abstract entities of mathematics; this bridge is provided by what he calls 'impure abstract entities', which are, for example, "functions that map physical objects into pure abstract entities"[64]. Hence, the mathematical theories that he considers "include at least a minimal amount of set theory with urelements (a urelement being a non-set which can be a member of sets)" and they "must also allow for non-mathematical vocabulary to appear in the comprehension axioms"[65] so that at the end they involve both the mathematical and the physical vocabulary together. After having established what a mathematical theory and what a nominalistic physical theory are, along with the bridge, the one-place predicates $M(x)$ meaning 'x is a mathematical entity' and $\neg M(x)$ meaning 'x is a non-mathematical entity', that is required to link the two, he states the following theorem that shows mathematics to be just the conservative extension of the physical theory and hence renders it dispensable.

Theorem 1 *(Principle C (for conservativeness))For A any nominalistically-stated assertion let A* be its corresponding restricted assertion in which each of its quantifiers has been restricted with the formula 'not$M(x_i)$', and for N any nominalistically-stated body of assertions let N* consist of all assertions A*; and let S be any mathematical theory. Then A* isn't a consequence of $N^* + S +' \exists x \neg M(x)'$ unless A is a consequence of N.*

---

64 Ibid.

65 Ibid.

Notice that the inclusion of the axiom '$\exists x \ulcorner M(x)$' is necessary so that the mathematical form of the physical theory is really a conservative extension of N. Without it, $N + S$ may be inconsistent since N as a nominalistic theory may rule out the existence of abstract entities. If we restrict each quantifier of the nominalistically-stated assertions A of N with the formula $\ulcorner M(x)$ and call the resulting ones A^* and N^* respectively, then N^* is an 'agnostic' version of N which allows for statements that may include both mathematical and non-mathematical entities. Hence, this formula allows for A^* statements like 'all non-mathematical objects obey Newton's laws' but at the same time it allows for the possibility that there may be mathematical objects that do not; this possibility does not exist in N.

The theorem above follows from the stronger theorem

Theorem 2 *(Principle C') Let A be a nominalistically-stateable assertion, and N any body of such assertions. Then if A^* is a consequence of $N^* + S$, it is a consequence of N^* alone. ($N^* + S \vdash A^* \Longrightarrow N^* \vdash A^*$)*

which in turn is equivalent to the following:

Theorem 3 *(Principle C") Let A be a nominalistically-stateable assertion. Then A^* isn't a consequence of S unless it is logically true.*

What follows then from these theorems is that A^* is not a consequence of $N^* + S +' \exists x \ulcorner M(x)$' unless A is a consequence of N and therefore, mathematics constitutes just the conservative extension of the theory. This is also known as the conservative extension theorem and lies at the heart of Field's argument.

To demonstrate in what way the mathematical fictions may be useful, Field examines arithmetic, geometry and distance.

2.2.3 Illustration of Why Mathematical Entities are Useful: Arithmetic, Geometry and Distance.

Using the examples of arithmetic and of geometry, Field shows how, by using mathematics, one can construct a conservative extension of these two, otherwise nominalistically formulated bodies of assertions. Then he uses these extensions -and therefore abstract, mathematical premises- to prove claims that rely only on the original nominalistic ones. In these proofs, which could be done even without using abstract mathematics, numerical claims are just abstract counterparts of purely arithmetical or geometrical claims and this indicates, according to Field, that they are actually not necessary, just useful and truth-preserving devices. But the fact that mathematics (or the theory of real numbers plus set theory) is truth preserving does not entail that it must be true as well. And therefore, he concludes, we only need to assume that it is conservative. Moreover, it is a rather restricted form of conservativeness that is actually needed, and this restricted form follows from the consistency of set theory alone.

At this point we would like to agree with Fields that there is no logical necessity indicating that mathematics must be true. Yet, we can see a "utility-necessity" of numbers and set theory if we want to make measurements of the kind we are used to in physics[66]. And measurement, we believe, is above all, what geometry and arithmetic are about -at

[66] Hilbert's representation theorem aknowledges a certain utility of real numbers in geometric reasoning and even Field agrees witht that. Given this utility and ignoring for the time being the weaknesses of both Hilbert's and Field's programmes one can see that numbers are usefull devices even if they are nothing more than just that.

least in their applications in physics. Just bear in mind that the word Geometry itself means precisely that: to measure the earth!

2.2.4 Nominalism and the Structure of Physical Space

So far, Field has tried to establish that numbers are not necessary in doing physics. Instead, he claims, the quantifiers that we need in order to derive what we want range over space-time points that do exist. From a Platonic point of view, our knowledge of mathematical structures is a priori, while our knowledge of the structure of physical space(time) is an empirical fact, subject to experientially-based revision. Moreover, the postulate of points of space is less rich than that of real numbers, for the simple reason that the operations of addition and multiplication that go together with the postulate of the real numbers, do not go with space; for we cannot define addition of two points, nor multiplication. The similarity in structure between space(time) points and mathematical objects should be of no surprise to anyone, Field claims, nor should be regarded it as an amazing coincidence, because all the mathematical artifacts, like real numbers, differentiation and so on were developed in response to certain theories developed in order to deal with space and time. As a result of this close connection, one should expect that these mathematical theories have strong structural similarities to the physical structure of space and time.

Based on this conviction, he claims that relationalist views of spacetime would be a violation of nominalism, as opposed to the substantivalist view[67], and that the problem

[67] According to the substantivalist view, space-time points and regions are entities that exist in their own right.

According to the relational view, spacetime is characterised in terms of physical objects -actual or possible- and it takes one of the two forms: reductive and eliminative relationalism. Reductive relationalism claims that space-time points and regions do not have a separate existence but they are some kind of set-theoretic

for relationalism is especially acute in the context of field theories. "If the field is defined as an assignment of some property to each spacetime point", he writes, "this assumes that there are spacetime points. So a relationalist would have to either avoid postulating fields or come up with some different way of describing them"[68]. Further, he nominalizes the Hilbert formulation of Euclidean geometry by allowing the first order variables to range over points or regions of the spacetime only, and both the first and the second order quantifiers to range only over regions of spacetime.

The issue Field raises here is very interesting and highly controversial for more than one reason and from more than one aspect. Shapiro and Malament have attacked both the view that using spacetime points instead of numbers makes a real difference and that substantivalism is necessarily the position to adopt about spacetime points, as we shall see shortly. We, on the other hand, will come back to the point he makes about fields in chapter four.

In a way analogous to the one already used to nominalize geometry, Field tries to nominalize physics as well. The principle he employs in this attempt is that "underlying every good *extrinsic* explanation there is an *intrinsic* explanation"[69], where by extrinsic he means explanations -or functions- that use certain extrinsic constant numbers, like for example the gravitational constant[70]. If the principle is correct, he claims, the real numbers

---

construction which is based on the physical objects and their parts. According to eliminative relationalism, on the other hand, it is illegitimate to quantify over unoccupied space-time points or regions, while quantification over occupied ones is fine since this could be regarded as equivalent to quantifying over the objects that occupy them. (Field, p.34 & 114)

68 Ibid., p.35.

69 Ibid., p.44.

70 Field associates the extrinsic 'quality' of a constant with the fact that as it is just a real number, it does not play any causal role in the forces acting between two bodies.

must be eliminated from physical explanations; and they have to be eliminated because otherwise the explanations are arbitrary and hence unsatisfactory. In the meantime, he does not exclude altogether the use of mathematics in scientific explanations, because they are truth preserving and as such they may be used as auxiliary devices in inferences; in this sense they are part of the extrinsic explanation and therefore they are dispensable. For that reason and considering the description of fields using tensors, Field does not like the arbitrariness of choosing units of distance, although he approves, of course, of the fact that we do not need to use numbers with them. But as we just said, we will return to his views about fields later on, when we have presented fibre bundles and the description of gauge theories through them.

2.2.5 A nominalistic Treatment of Newtonian Gravitational Theory

Briefly, we outline in this section Field's strategy to nominalize Newtonian spacetime, and we do so because we need the main ideas in order to understand the criticisms against his proposal. Field considers that we need three axioms in order to account for betweenness, congruence and simultaneity, and these are his primitives. All other genuine spacetime relations, he believes, are defined in terms of them. Using the example of temperature as a typical physical quantity, he introduces temperature-betweenness, temperature-congruence and temperature-less relations among spacetime points -rather than introducing betweenness and congruence among temperature properties. Doing it this way, he claims, one gets the desired representation and uniqueness theorems that are necessary in a theory of measurement. Having done that, he defines any scalar primitives, like the gravitational potential

and the mass density in this specific case, in the same way that he does with temperature. Then he is able to introduce a joint axiom system, what he calls JAS[71], which contains all the necessary primitives and nothing more. All these are defined on the same set of space-time points and thus he creates a working model. For each such system with appropriate axioms there is both a spatiotemporal function φ from spacetime onto $\mathbb{R}^4$ and a scalar representation function ψ also from spacetime and onto an interval of the real numbers. Each of these functions is unique up to the appropriate class of transformations. The physical laws are usually expressed as functions $T = \psi \circ \varphi^{-1}$ mapping quadruples of real numbers into real numbers (a one-to-one map) and they express the interrelation between the two functions. So, laws about T can be restated as laws about the interrelation of φ and ψ and vice versa; and since the two functions can be restated in terms of the axioms of the JAS, so can their interrelations.

Let us come back to a point we raised before. It seems that what Field fails to recognize here is that the numbers are there to represent measurable properties of the physical entities. Without the numbers, or something like them, there is no chance of relating theories to physical world. Choosing spacetime points as the real and truly existing entities, we just require an extra, intermediate step when measuring, say, distance between two such points. And although a device like that allows for measurement, we tend to believe that the measurability should be intrinsic to any good theory, because we believe that a scientific theory is an intertwined combination of mathematics, interpretations and connections with the world, where connections are the experiment and the measurement.

---

[71] p.59

2.2.6 Criticism of Field's programme by Malament

Field's programme, though very imposing and ambitious, has been criticized ever since it appeared. Malament, in his 1982 review of the book criticizes the programme from three different perspectives. Using the example of the Klein-Gordon field and calling T a nominalist reformulation of the theory, in other words a set of sentences in an appropriate nominalist language L (a second-order language with variables for individuals as well as the sets of individuals and the relation symbols '=', 'Seg-Cong', 'Scale-Bet', 'Scale-Cong', '$\in$ ', '$\leq$'), S some fact about the field and S_L its nominalist reformulation in L, Malament claims that in order for T to rebut the indispensability argument at least three conditions must be met:

1. L qualifies as a nominalist language.

2. All assertions concerning the space-time distance function and the Klein-Gordon field which are essential for the purposes of science can be reformulated in L.

3. Given any sentence in L, if it is derivable from the theory of the Klein-Gordon field in its original formulation, then it is a logical consequence of T.

Condition (3) is guaranteed by the representation theorem, claims Malament, but in his view the other two are highly controversial. Condition (2) restricts the claims that can be made far too much. If we accept that the Klein-Gordon field determines a set of models of the form $((M, d), \psi)$, where (M, d) is a Minkowski spacetime and ψ is a smooth real-valued function on M satisfying the Klein-Gordon equation, there are three different types of theorems about this set:

A. Propositions which report generic features of individual models.

B. Propositions that establish the existence of models with special features.

C. Propositions that make essential reference to more than one model.

At best, Field can reformulate in his language only theorems in the first category because even if he enriched his language L to allow reference to other qualitative relations apart from congruence and betweenness, "he cannot do anything except assert general truths about what goes on within arbitrary models". In other words, he cannot establish the existence of, say, a Klein-Gordon field that is non-constant nor can he establish that two models $((M, d), \psi)$ and $((M, d), \psi')$ may be deterministically linked. The reason he cannot do the first is that although he can define non-congruence between spacetime points -and hence between fields defined over these points- the statement can only capture the fact that the Klein-Gordon field is non-constant in all cases. As for the second, determinism involves 'agreement' between the two models on a simultaneity slice H in (M, d) such that "if ψ and ψ' agree on H and if their time derivatives (i.e. directional derivatives orthogonal to the slice) agree there, then ψ and ψ' agree everywhere". What Malament calls 'agreement' is a direct relation between ψ and ψ' and it is a lot richer than can be captured by congruence and betweenness.

So far as condition (1) is concerned, the language L needed for the description of something as complex as the Klein-Gordon field, or Hamiltonian mechanics or quantum mechanics, is too rich for nominalism, Malament claims and we entirely agree. For one reason, in the case of the Klein-Gordon field, the language admits second order quantifiers, quantifying over both spacetime points and sets of spacetime points. Field disputes this point claiming that the quantifiers range over *regions of spacetime points* rather than *sets*,

even though he recognizes that the character of logical-consequence in L is thus rendered problematic: second order logic is not complete, i.e. it is not recursively axiomatizable and this he would rather avoid. Field conjectures that one might be able to do physics with a weakened, first-order version; however, if this conjecture fails, one would rather keep the logical resources of the nominalistic language than abandon nominalism altogether. But this is exactly where the problem lies, as Malament points out, because "the logical consequence relation cannot be recovered in terms of a formal derivation system"[72]. Moreover, even if we brushed aside those problems, it is hard to see how a nominalist could justify the quantification over either spacetime points or spacetime regions. Though Field attempts to justify this by asserting that the substantivalist view of spacetime is the correct one, Malament finds the response unsatisfactory, and that not only because the controversy between substantivalists and relationalists is not conclusive. Rather, it is the claim that spacetime points, unlike abstract mathematical objects, are concrete entities that exist in their own right to which he objects. As he put it: "But I, for one, begin to lose my grip on the distinction when thinking about such things as 'spacetime points'. It would have helped me to understand his conception of nominalism if Field had explained how he draws the line and made clear why spacetime points are so much *better* than, for example, sets and qualities. If what constitutes a nominalistic language in the case of the Klein-Gordon field is hard to pin down, then things become completely out of hand in classical Hamiltonian mechanics and in nonrelativistic quantum mechanics. In the first case one would have to quantify over possible dynamical state, while in the second even if they could think of the theory as deter-

[72] At this point Malament anticipates Shapiro's criticism and the application of Godel's incpmletness theorem.

mining a set of model -each a Hilbert space- one would not be able to find a representation theorem".

The issues Malament raises are very important and directly related to how Field's programme might (not) be applied in field theories in general, but to this we will come back in chapter three. In the mean time, we will examine Shapiro's objections to Field's nominalization programme, which are based on the problems that arise from Malament's condition (3).

2.2.7 Criticism of Field's programme by Shapiro

According to Shapiro, Field's programme for the development of a synthetic science fails for a similar reason that Hilbert's finitary mathematics fails as well. Since the two programmes are structurally analogous, the same criticism applies to both, and hence they both falter over Gödel's incompleteness theorem. More specifically, Shapiro shows in his paper *Conservativeness and Incompleteness* that there is an ambiguity in the formulation of conservativeness "which involves the distinction between semantic consequence and deductive consequence", a distinction that Field himself pointed out too[73]. Field's nominalistic physics is formulated in second order, as we have seen, whose first order variables range over spacetime points while its second order range over spacetime regions -rather than sets of points, although in the last chapter of his book he proposes a nominalization using just first-order language. These formal theories have to be recursively axiomatizable

[73] There are two senses of consequence (or implication) in logic, the syntactic, usually denoted by the simple turnstile $\vdash$ and the semantic, usually denoted by the double turnstile $\models$. The syntactic consequence $\Delta \vdash \phi$ suggests that ϕ can be proved formally from (axioms) in Δ, while the syntactic consequence $\Delta \models \phi$ suggests that ϕ is true in every model of Δ.

and complete. However, second order theories are known to be incomplete, since Gödel's completeness theorem does not hold for them[74]. This means that in theories such as Field's nominalistic N and extended $N + S$ "conservativeness is ambiguous as to whether it involves proof-theoretic derivability in N and $N + S$ or semantic consequence in N and $N + S$". Field himself has established only the semantic conservativeness, but this is not enough to guarantee that S is just the conservative extension of N. Shapiro, as a matter of fact, provides a counterexample that refutes the deductive conservativeness of S over N, by finding a sentence θ formulated in the language of N such that $S + N \vdash \theta$ but $S \nvdash \theta$, and he points out that "given semantic conservativeness, θ is true in all models of N but it is not deducible in N". Hence, for second order theories he shows that deductive conservativeness is not coextensive with semantic conservativeness.

If one tried to stick to first-order version of nominalistic theories, on the other hand, one cannot prove the existence of homomorphisms from spacetime points to R^4, which was another necessary requirement for the formulation of nominalistic physics, even though one maintains deductive conservativeness. Hence, either way, Field's programme runs into apparently insurmountable problems.

The question that arises, then, is that if Field's programme runs into such difficulties, why should we get into the trouble of discussing it? Despite the flaws of the programme, Field's denial of the indispensability of mathematics is an idea that is worth investigating.

---

[74] The theorem could be stated as follows:

$$\Delta \vdash_2 \phi \Rightarrow \Delta \vDash_2 \phi$$

but it is possible that

$$\Delta \vDash_2 \phi \text{ and } \Delta \nvdash_2 \phi.$$

where $\vdash_2$stands for provable and $\vDash_2$for semantic consequence, while the subindex 2 indicates second order logic.

Putting the logical arguments aside, or maybe along with them, we will see later on in this thesis that within the context of quantum field theories what Field would consider as purely mathematical structure -and hence dispensable- is essential and it contains vital information about the physical systems that the rest of the theory -its physical part- does not.

2.3 Structuralism

The main philosophical idea behind structuralism is that the *essence* of *mathematical objects* is their relations to other mathematical objects and the *structures*[75] in which they are arranged. Mathematical objects, like the natural numbers for example, are ontologically dependent in the sense that they only exist -if at all[76]- in relation to other natural numbers. As Shapiro put it: "The subject-matter of arithmetic is a single abstract structure, the pattern common to any infinite collection of objects that has a successor relation, the unique initial object, and satisfies the induction principle. The number 2 is no more or less than the second position in the natural number structure; and 6 is the sixth position. Neither of them has any independence from the structure in which they are positions, and as positions in this structure, neither number is independent of the other"[77]. And according to Resnik, another proponent of structuralism, natural numbers "have no identity or features outside a structure", so they must be regarded as "structureless points or positions in struc-

---

[75] Resnick, in his 1997, declares a preference for the term 'pattern' rather than 'structure', because, as he puts it, he finds "it more suggestive to speak of mathematical patterns and their positions, rather than structure" (p.202).

[76] Strucutralists' views over the existence of mathematical objects differ. So, Shapiro and Resnik, for example, are realists in ontology, while Benacerraf and Hellman are realists in truth value only.

[77] Shapiro, *Thinking about Mathematics*, p.258.

tures"[78]. Hence, unlike the ontological Platonist, who could say that mathematical objects, like the numbers, are ontologically independent from each other -just like a physical object is ontologically independent form another physical object- a structuralist would insist that such objects are not ontologically independent, because the essence of their existence is their relations to other objects of the structure they belong to and hence they are nothing other than places within the structures. Yet, numbers are epistemically independent since one may know about a specific number -say 8- while at the same time may not know about another -for example 1786.

Shapiro defines[79] "a *system* to be a collection of objects with certain relations among them" and "a *pattern* or *structure* to be the abstract form of a system, highlighting the interrelationships among the objects and ignoring any features of them that do not affect how they relate to other objects in the system". Then, he claims, we understand structures via a process of abstraction, where we focus on the relations between the objects. Obviously, more than one system may exemplify one structure, hence a structure is one-over-many. As Shapiro points out, "the traditional examplar of one-over-many is a *property*, sometimes called an *attribute*, a *universal* or a *Form*"[80]. Hence, from the structuralist's point of view, "a system is a collection of objects with some relations between them and a structure is the form of a system"[81].

A Platonic view of universals, known as *ante rem realism*, advocates that the existence of some universals is independent of whether their instances exist or not, hence, the

78 Resnik, 1981.

79 *Thinking about Mathematics,* p.259.

80 Ibid. p.262.

81 Ibid.

'one-over-many' comes first, while the 'many' comes second. Contrary to this view and in accordance with the Aristotelian *in re realism*, the universals are nothing more or less than their instances, in which case the 'many' is prior to the 'one-over-many'. The conceptualists believe that the universals are mental constructions, while the traditional nominalists either consider them as non-existent or think of them as linguistic constructions.

Although the discussion about what we mean when we say that structures exist independent of the systems or objects that exemplify them, and about how we get to know them, is long, for the purpose of this thesis suffice it to say that we will be considering the structures in an ante rem sense, that is to say, as existing prior to and independent from their instances. As for the epistemological question, it will do to say that we get to know these structures via pattern recognition and through abstractions.

2.4 Michael Redhead's Surplus Structure

Michael Redhead (2001) claims that the relation between physics and mathematics is of a structural character. He talks about two different types of structure, a mathematical structure M and a physical structure P both of which could be regarded as models for an uninterpreted calculus C. The mathematical structure M may be considered as consisting of *isomorphism classes* of concrete mathematical structures "where two concrete structures in the same isomorphism class are related by a bijective correspondence which preserves its system of relations in the sense that if in the one structure the elements $x_1x_2, ...x_n$ satisfy the n-ary relation R, then the corresponding elements $y_1, y_2, ..., y_n$ in the second structure satisfy $R'(y_1, y_2, ..., y_n)$ if and only if $R(\ x_1x_2, ...x_n)$, where R' is the n-ary relation in the

second structure that corresponds to R in the first structure"[82]. The abstract structure, then, may be considered to be the universal or form that is shared by all the concrete structures in an isomorphism class. This second-order abstract structure, he claims, is what is associated with physical reality and it "can be thought of as a second-order property of the 'true relations' rather than the true relations themselves"[83]. This notion of abstract structure, Redhead points out, dates back to the early writings of the empiricist tradition[84].

As to the question of what exactly is a concrete mathematical structure, Redhead acknowledges that the question is formally problematic[85] and distinguishes mathematical structures which are specified categorically in an intuitive Platonic sense. Thus, Redhead's concrete mathematical structures belong to a unique isomorphism class and are different from Shapiro's algebraic structures which involve many isomorphism classes.

Redhead believes that this kind of concrete abstract structure reveals to us the relation between mathematics and physics, since an abstract structure is associated with a physical system as well as with a mathematical structure; hence, structures involving the natural or the real numbers, may belong to the same isomorphism class that maps a specific physical structure onto each of these mathematical structures. Therefore, a mathematical structure can be used to represent a physical structure.

---

82 M. Redhead, 2001.

83 Ibid.

84 See, for example, Russell (1927) or Carnap (1929).

85 The problem, as we have seen, originates from the fact that second order logic is not complete, while first order logic which is comlete cannot provide categorical models.

To show how mathematical structures are used to represent physical structures in the context of measurement, he uses -among other- the examples of temperature and mass as they are measured by natural numbers. He writes:

> Consider the case of ratio scales of extensive quantities, such as mass. Such quantities map onto a one-dimensional vector space spanned, for example, by the unit of mass. Given the choice of unit (base vector), the *measure*, i.e. the ratio between the quantity and the unit, is specified by a dimensionless number which represents the physical mass relative to the choice of unit. But again, the representation is not unique. Changing the unit by a factor α rescales the measure by a factor α^{-1}.
> Another very familiar example of the underdetermination of mathematical representations is the variety in the choice of coordinate maps or charts for the (local) representation of a physical manifold, such as the phase space in mechanics or the space-time manifold. The choice of chart is a matter of convention, and is to be decided by pragmatic considerations of convenience, simplicity and so on.
> Or, as a final example consider interval scales such as are used to measure temperature. Both the unit and the zero of the scale, are arbitrary and hence the numerical representation is unique only up to a linear transformation. For example, consider the conversion of temperature T_c of the Centigrade scale to T_f of the Fahrenheit scale by the transformation $T_c = 5/9(T_f - 32)$.

One thing that becomes apparent from the above examples is that the choice of the mathematical concrete structure that represents a given physical structure is not unique. There is no necessity whatsoever to dictate that only one out of the many mathematical structures which belong to the same isomorphism class with the physical structure is its 'correct' representative.

Field in his programme tries to avoid this problem[86] by getting rid of all arbitrary constants (conventions as he calls them) together with all the other numbers. In Shapiro, on the other hand, all the members of a structure share the same relations, so the different

[86] From Field's nominalistic point of view, the use of any numbers, constant or not, is forbiden -numbers should play no essential role in science. The structural underdetermination, however, involves the use of constants for conversions from one scale to another and this should be avoided if one wanted a unique representation of physics by nominalistic mathematics.

representations are not essentially different since they exemplify the same bunch of properties. The members of Redhead's isomorphism class share the same relations too, through bijective correspondence, something that makes it also into a many-over-one. So, in an isomorphism class the n-ary relations obtaining in one structure in the class correspond to n-ary relations that are shared by the objects of the other structures in it. In this manner, the ambiguity of representation of this type is a consequence of the fact that there are more than one concrete mathematical structures isomorphic to a given physical structure. Schematically this may be represented as follows.

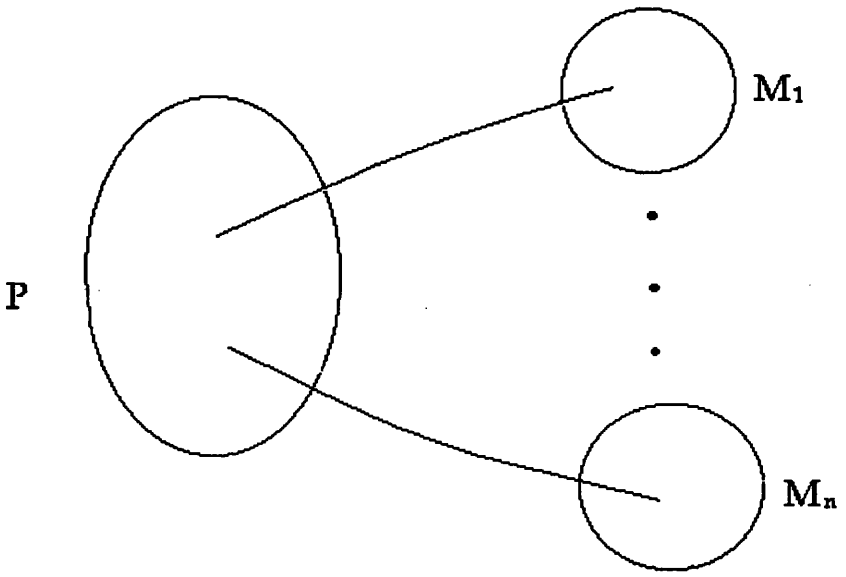


Figure 1
Ambiguity of Representation of the First Type

The fact that the choice of a representative for a physical structure is *decisively* conventional and, therefore non-unique, raises the question: has the conventional choice of

mathematical representation of a physical system got to do anything with physics? The answer to this question will come after we have considered ambiguity of representation of a different kind, one which is related to the notion of symmetry.

Apart from the aforementioned ambiguity, which we will call it ambiguity of the first type, or *ambiguity of which mathematical structure to choose* and which is, as we have already mentioned, the end result of having too many concrete mathematical structures in the isomorphism class which includes the physical structure that we are aiming to represent, in physics we have two more types of ambiguity. The second one, which we call *ambiguity within the same structure*, is related to the notion of symmetry, which in turn are related to conservations of physical quantities. The third type is also related to the notion of symmetry and has considerable physical import, since, as we shall see, certain symmetries of physical systems are related not only to conservations of physical quantities but also to interactions. But, first things first, we need to examine the notion of symmetry.

2.4.1 Symmetries

Using the map-terminology, symmetries are expressed as bijective structure-preserving maps of a structure onto itself -an automorphism of the structure. This kind of symmetry, is related to ambiguity of representation within a given mathematical structure where the same object of P, the physical structure, can be mapped on two different objects of the same mathematical structure M through two different isomorphisms $x : P \rightarrow M$ and $y : P \rightarrow M$. Then, $y^{-1} \circ x : P \rightarrow P$ is an automorphism of P and $y \circ x^{-1} : M \rightarrow M$ is an automorphism of M. In spacetime models, the automorphism $y^{-1} \circ x : P \rightarrow P$ of

P is referred to as a point transformation or as an active symmetry of P[87], while the map $y \circ x^{-1} : M \rightarrow M$ is known as coordinate transformation or passive symmetry of M[88].

The symmetries of the physical system P express important structural properties of P. A structure, as we have seen, is a collection of objects with their relations. A symmetry within a physical system expresses the fact that two distinct parts of this structure can be mapped onto each other, or the fact that these two parts are indistinguishable with respect to certain properties. So, take for example a physical system that contains all objects which interact according to Newton's laws of motion and the universal law of gravity, along with their classical -i.e. non-relativistic- spatiotemporal relations. Consider within this structure a system S comprising two bodies with mass m_i and m_j respectively that occupy some spacetime region r. Using the automorphism $y^{-1} \circ x : P \rightarrow P$, map this system onto another distinct system S', which contains two bodies with masses $m_i' = m_i$ and $m_j' = m_j$ in region r'. The automorphism is preserving the relations inside the two systems, which means not only that the two bodies in each system obey the same laws but also that the exact values of -say- their velocities and relative positions are the same. This, in turn, indicates that within this structure the space points are indistinguishable or that space is homogeneous and isotropic -in other words, the background gravitational field is constant. So, in this case the invariance under space translations and rotations reveals homogeneity and isotropy, which is a structural property of P indeed. In this case, the ambiguity of representation is expressed through the automorphism $y \circ x^{-1} : M \rightarrow M$

---

87 Active because it maps one object of the physical structure into a different ofbject.

88 Passive transformation because it maps one coordinate system onto another. This transformation takes place within the mathematical structure and does not involve any transformations of the physical structure.

which basically reflects precisely this structural property of P and is backed by the fact that the two structures have the same property, that is to say, the space vectors which represent coordinate systems are invariant under rotations and translations.

With these examples it has become clearer, we believe, that ambiguity in the representation of physics by mathematics is inevitably there, but we need to clarify how ambiguity within the same structure has physical significance. Discussion of this second point will become clear after we introduce the notion of surplus structure.

2.4.2 Surplus Structure and Gauges

In many cases in physics, the mathematical structure M that is isomorphic to the physical structure P is a substructure of a wider structure M'. This basically means that the objects and the relations between them that can be found in P have corresponding objects and relations in M alone. The rest of the structure M' is what Redhead calls *surplus structure* and it includes objects as well as relations both between elements of the surplus structure only and between them and the objects of M. Hence, the surplus structure is a structure indeed and not just a set of (excessive) elements.

There are several examples from the practice of physics where this happens. For example, Redhead mentions (2001) the use of complex currents and impedances in alternating current theory and the S-matrix theory of elementary particles scattering which makes use of the complex plane. In both cases, the physical system is mapped onto the real part of the complex plane, yet the use of complex numbers facilitates calculations and derivations. Other examples, like that of the total energy, i.e. the sum of kinetic and potential, of a me-

chanical system, illustrate how some entities initially believed to be members of the surplus structure eventually became part of the physical structure itself.

One such case where some quantity from the surplus structure tries to break into the physical system itself is the case of electromagnetism, which is a theory with a gauge symmetry, and the quantity is nothing other than the gauge potential, which we usually denote by A_μ. In chapters three and four we will examine in detail different ways of interpreting this kind of symmetry and attributing physical significance to the terms that appear there. For now though we will only attempt a comparison between the nominalist and the structuralist views using the three examples we have just mentioned, plus a fourth one that will help us illustrate the differences and the similarities between them.

This other physical system that will be of interest to us in the following chapters is one which contains all objects that carry electric charge. Take a system S_1 with two of these objects of P with charges e_i and e_j respectively. An automorphism $y^{-1} \circ x : P \rightarrow P$ would be one that takes S_1 onto another system S_2 with same charges and electromagnetic fields. In the mathematical structure, there is one more element associated with the electromagnetic fields, the potential -or gauge field- and there is a freedom as to which gauge we may pick for any specific electromagnetic field. In other words, there is a symmetry present in the mathematical structure. The objects in the two systems have the same equations of motion and the mapping preserves their relations. But the presence of the potentials gives rise to coupling terms that allow for a description of the interactions between them, as we shall see in detail in the following chapter. The very fact that they *interact* can be considered to be a structural property of the structure P where they belong and the invariance of the elec-

tromagnetic field under a local change of the gauge can be considered as an expression of *this property*. In the mathematical structure that belongs to the same isomorphism class as the objects with charge, this structural property is expressed by what we call covariance of its objects under the local group of $U(1)$ transformations and since the two are structurally the same, we can conclude that a mathematical structure with covariance under local $U(1)$ transformations allows for description of electromagnetic interactions, albeit this description comes from what we call the surplus structure, that is, the part of the mathematical structure that does not have a physical counterpart[89].

So, once again, ambiguity in the representation, but this time within a given mathematical structure, gives away/reveals/describes physical relations or, in this case, interactions that are a structural property of P. The difference, though, between what we have described here and the symmetry example of the previous section is that in the previous case change might occur in either the physical system or the mathematical structure, while in the case we are considering here change occurs in the mathematical surplus structure only and this in a sense controls the physical system since it allows for the description of interactions that take place in it. Hence, although this third type of ambiguity resembles that of the second type in the sense that both relate to symmetries present in the structures and both give rise to conservation laws, the third type of ambiguity occurs in structures with symmetry transformations that do not affect -that is to say, do not actively change- the physical system nor the objects in it but they give rise to coupling terms that, as we shall

[89] One might object at this point that had we adopted a realist view of the gauge field and an active interpretation of the gauge transformation, the different gauge fields could be considered as different physical entities. However, anticipating the arguments that will unfold in chapter 4, we are assuming here that this option is not viable.

see, are usually interpreted as interactions; hence we will call it *ambiguity that gives rise to couplings*. Although a drawing will not do justice to what is really happening in this rather complicated case of symmetry, a very schematic way of representing it is attempted in the following diagram, where we have just depicted symmetry transformations in the surplus structure.

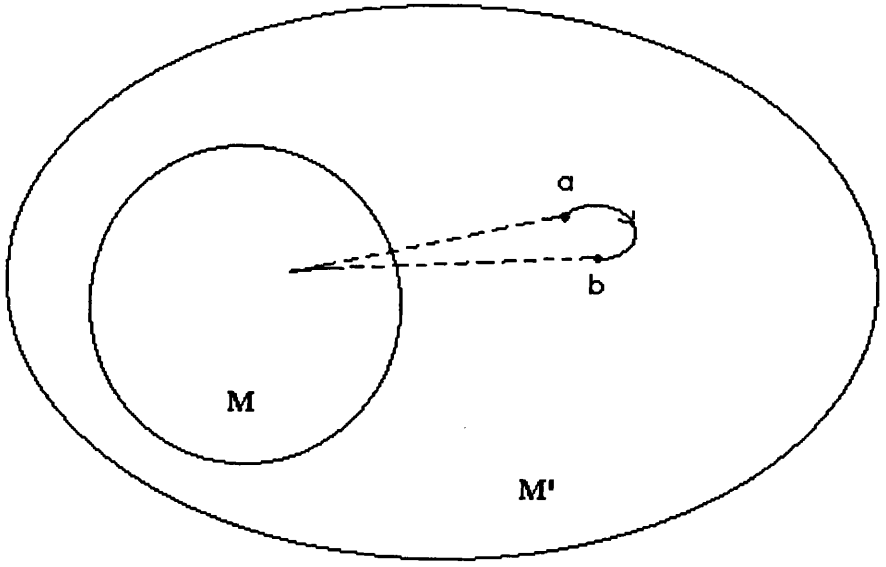
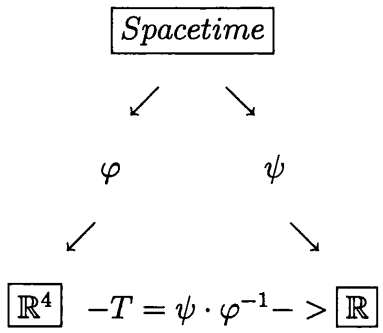


Figure 2
Symmetry Transformation of Surplus Structure

Before we conclude this section, let us try to clarify how one may understand this illustration of symmetry transformations in the surplus structure. M stands for the mathematical structure, while M' corresponds to the surplus structure. There may be maps that take you from M to M' and back, which map one element of the structure to a bunch of elements in the surplus structure. The images are equivalent in the sense that correspond to the same entity of M. Certain elements of M' that describe the transformations between a and b are also used for the description of interactions between the structures of M.

2.4.3 Comparing Field & Redhead

At first sight, Field's and Redhead's approaches seem very similar. One question that arises then is how is M. Redhead's surplus structure different from Field's conservative extension? In order to point out the differences, let us indicate the similarities between the two approaches first. Consider the example of the temperature we mentioned above. Field, in this case, would say that for a nominalistic account we take the spacetime points to be primitive objects and we define temperature-congruence and temperature-betweenness relations between spacetime points, thus defining the scalar primitive temperature. This combination has been given the name JAS. Then, "for any model of the combined system there is both a 1-1 spatiotemporal function φ onto $\mathbb{R}^4$ and a scalar-representation function ψ onto an interval, each function unique up to (but only up to) the appropriate class of transformations. Now, physical laws governing a scalar like temperature are often expressed as laws about a scalar function $T = \psi \cdot \varphi^{-1}$ mapping quadruples of real numbers into real numbers"[90]. Therefore, laws about T could be expressed as laws about φ and ψ alone, while we can always go to $\mathbb{R}^4$ or to $\mathbb{R}$ to calculate, derive and so on, whenever this is necessary. Schematically this could be represented as follows:



[90] Field, p.59.

For Redhead, on the other hand, physical bodies with the property 'temperature' constitute a physical structure P, which belongs to the same isomorphism class as the structure of the real numbers M. The two share the same properties and through a mapping $\mathcal{T}$ we can go back and forth. Schematically:

$$\boxed{P} < - - \mathcal{T} - - > \boxed{M}$$

In this case, one could claim that basically the two are similar if we considered Field's JAS and $\mathbb{R}^4$ to be the same as Redhead's P and $\mathcal{T}$ to be the same as $T = \psi \cdot \varphi^{-1}$. Despite the similarity of the two approaches in this simple case, if we go to a more complicated and richer physical structure, like the one of objects which interact electromagnetically, for example, the differences between the two approaches become manifest; let us see how. Redhead would claim that in the case of electromagnetism, the physical structure contains charged particles, electromagnetic fields and their relations. The mathematical structure would be that of $U(1)$, along with the surplus structure it involves, and that the gauge field which allows for the interactions in P belongs in this surplus structure (for the moment we leave aside, once again, the controversy of whether the gauge field A_μ is indeed part of the surplus structure or not and just take it to belong there). Once again, the two structures belong to the same isomorphism class and they are related through a relation preserving map. The surplus structure, as we saw above, allows for the description of interactions between the physical objects in P and the gauge potential A_μ plays a crucial role, as we shall see in chapter four. From Field's perspective, one could claim that it is possible to nominalize electromagnetism and its interactions by using a similar, though inevitably more compli-

cated, approach as before. Then again, as we see it, Field's nominalistic programme faces a major difficulty here. His primitives are spacetime points with scalar or vector relations among them. In the charge-free case his programme is better off because there one could consider the electromagnetic field as relations of a vector character between the spacetime points, and then using appropriate maps onto $\mathbb{R}^4$ and R one could get all the laws that govern it. However, in the case of electromagnetic field with charges, one might be able to get the relations between charges -the sources- and fields -their effects- only if one considered both the charges and the electromagnetic field as primitive relations among spacetime and presupposed that they are related to each other via the already known equations of motion; at least that is how he did in the case of Newtonian gravity. But then he would be hard pressed to also introduce the gauge potential as a primitive relation as well in order to account for the effects on electrons passing from areas where the actual electromagnetic field is zero whereas the A_μ field is not, and this field does not correspond to a physical quantity. Moreover, if the gauge field makes the transition from the surplus structure into the physical structure P, Field's approach will be proved unable to accommodate it.

On a more fundamental level, Redhead and Field differ in the following. Redhead considers the physical structure to consist of concrete objects and the theory to consist of all true statements about these objects. The statements of the theory, as he perceives it, are closed under deduction or in other words the theory is complete, but he does not assume the theory to be axiomatizable. He understands that in a rather intuitive Platonic sense and although he recognizes the problem of incompleteness of a second order formulation he does not attempt to offer any solutions to it. Field, on the other hand, begins assuming that

the theory is axiomatizable and hence he runs into the problem of incompleteness, that does not allow him to prove that the mathematical part of it is just the conservative extension.

The failure of Field's programme, as Shapiro showed, was due to the existence of a counterexample which was part of what he called the nominalistic assertions N without being a provable theorem in the original system. Contrary to that, the counterexample was derivable from the theoretical structure S alone. In chapter 4, we will attempt to show that in the context of gauge theories such a counterexample in fact exists.

Chapter 3
Formulations of Gauge Symmetries

3.1 Ambiguity of Representation of the Second Type and the Third Type: More Canonical Variables/Degrees of Freedom than the Ones Needed?

The aim of contemporary theoretical physics is to describe physical systems that interact[91] -after all, it is through interactions that we 'observe' physical phenomena in general and, in particular, phenomena that occur at very small scales[92] involving the so called elementary particles and nature's fundamental forces. These particular types of interactive systems have been successfully described using quantum mechanics and the notion of symmetry, which plays a crucial role as we shall see shortly.

In the previous chapter, we mentioned that the ambiguity of representation of the second type is related to notions of symmetry and symmetry transformations that may be considered to be active, i.e. transformations of the physical structure, or passive, that is mappings of the mathematical structure onto itself such that they do not correspond to any change of the physical system.

A very general idea of what a symmetry transformation is may be captured by the following simple illustration.

---

[91] Here we are referring to high energy theoretical physics that deals with elementary particles -or should we say fields?- and fundamental forces.

[92] Examples of what we mean when we are reffering to small scales: size of nucleous $\simeq 10^{-14}m$, size of quarks $\simeq 10^{-18}m$.

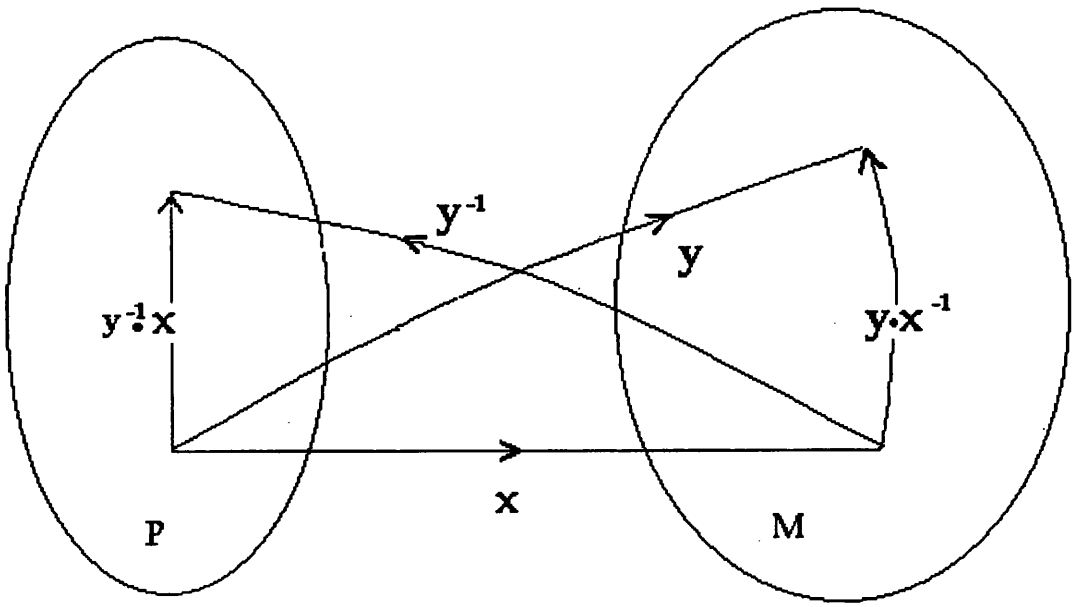


Figure 3
Ambiguity of Representation of the Second Type

P stands for the physical structure and M for a single mathematical structure that represents P. Between P and M there are more than one distinct isomorphisms -here depicted by x and y- that illustrate the ambiguity of representation of the second type. Associated with these two isomorphisms are automorphisms in both P and M - $y^{-1} \circ x$: $P \rightarrow P$ and $y \circ x^{-1} : M \rightarrow M$ respectively, that map elements of each structure onto elements of the structure itself. These automorphisms represent what we call symmetry transformations and they are considered to be active when they take place in P (i.e. $y^{-1} \cdot x$) and passive when in M (i.e.$y \cdot x^{-1}$).

The presence of symmetries in the mathematical representation -or description- of physical systems often manifests itself with the presence of more canonical coordinates -or degrees of freedom- than the ones necessary for the description of the physical system. This results in excessive mathematical structure which constitutes, as we have seen, an example of what Redhead calls the *surplus structure* in the mathematical representation of the physical system. Symmetry transformations affecting just the elements of the surplus

structure, but reducing to the identity on those mathematical elements directly correlated with the elements of the physical system are illustrated schematically bellow.

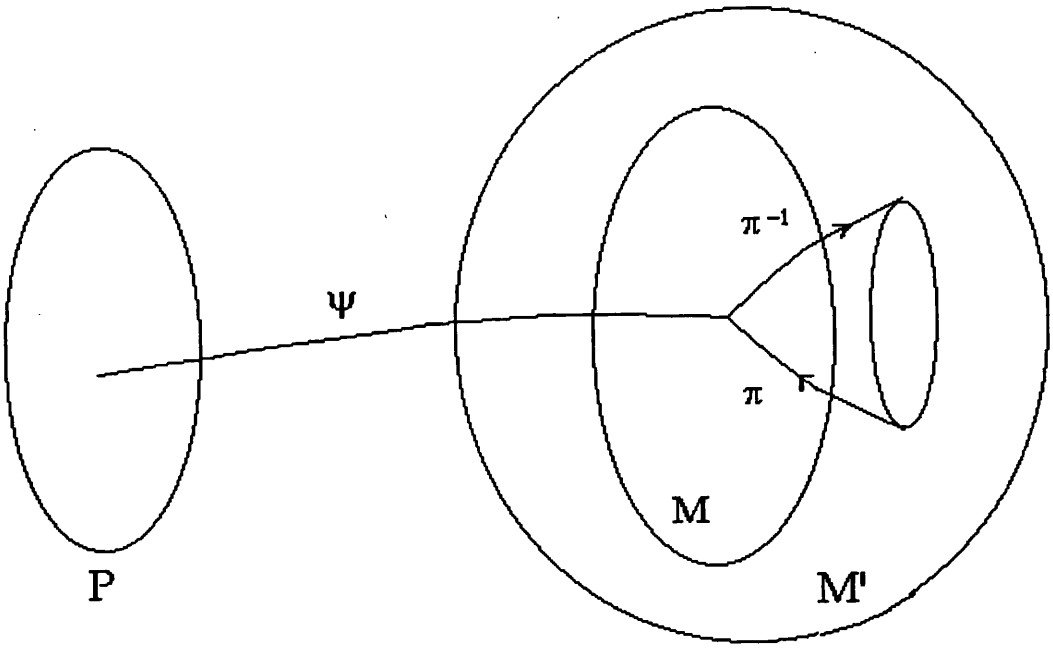


Figure 4
Ambiguity of Representation of the Third Type

This is the situation that arises in the case of gauge symmetries as we shall see in a moment and this is the case of what we are referring to as ambiguity of the third type.

In everyday manner of speaking, when one uses the term gauge one refers to either the measure or the unit of a quantity. If we generalized this notion[93] we could consider that the mathematical representation of any physical structure is the gauge for that structure. Ambiguity of representation of the first two types involves different unit-gauges, while ambiguity of representation of the third type results in different measure-gauges. In physics, however, we are used to narrower definitions of the notion of gauge. Leaving uses as in pressure gauge aside, we will focus on the notion of gauge as this became known in modern

[93] As Redhead in 2001.

theoretical physics, where it is inseparably connected to the notion of symmetry involving surplus structure. This type of symmetry, the gauge symmetry, has as a main characteristic the invariance of the theory under phase transformations. As it turns out, mathematical structures with gauge symmetries have been proved to be the most fruitful ones for the description of interactive fields, the elementary entities of nature, some would claim. In the course of their development, gauge theories have been formulated in various different ways but the ones that most often occur in the literature are the following. Gauge theories may be described as *constrained Hamiltonian systems*, a description involving presymplectic manifolds and in the philosophical literature it is favored by Belot (1996, 1998) and Earman (2000). In this description, gauge transformations are viewed as symmetries of constraints and are held responsible for the indeterministic nature of the first gauge theory to be examined, namely electromagnetism. The formulation in the form of *Yang-Mills gauge theories* introduces interaction fields in order to maintain covariance of the theory under phase transformations; these will turn out to be connections on a principle fibre bundle. This second formulation is favored by Lyre who uses the notion of *gauge freedom*, as a much more general notion than that of gauge transformation. However, the most general formulation of gauge theories is provided by the *fibre bundle formalism*, which features the advantages of all the aforementioned descriptions, plus a lot more as we shall see.

In what follows, we will present the formalisms with the intention of clarifying the role of symmetries in the representation of interactive physical structures and of raising the philosophical issues involved. But before we proceed, let us conclude this section with a brief comment on the notion of gauge. As we just mentioned, according to Red-

head (2001), gauge may be considered to be a mathematical representation of any physical structure, while gauge freedom is the ambiguity of either the first or the second or the third type involved in it. Hence, the notion of gauge freedom thus put forward accommodates all the three types of ambiguity. The problem here is that by giving the notion of gauge such a big scope one loses contact with the theory Weyl initiated; on the other hand, one re-introduces the original meaning of the word gauge, at least so far as ambiguity of the first type is concerned. Regardless of the advantages the more general use of the term may have, in this thesis we will be using the term gauge in its narrower sense, which is related to ambiguity of representation of the third type.

3.2 Gauge Symmetries and Constrained Hamiltonian Systems or Structures

The main purpose of this chapter is to set the framework in which gauge theories were first formulated and flourished. At present, all dynamic physical systems are described using variational calculus. There are two different approaches to the description of mechanical systems. One would begin with the equations of motion of the systems one is examining and then obtain the variational principle as a theorem, or, alternatively, one would assume the variational principle and derive the Hamilton-Jacobi or the Euler-Lagrange equations as theorems. So far there have been no indication that one of the two approaches is preferable to the other. There seems to be no physical necessity endorsing the second and as for the first, although the equations of motion entail the variational principle, there is no logical

necessity involved in that either. The belief that nature always acts in the simplest way, a belief shared by many, remains just a metaphysical predilection.

As it is well known, in non-relativistic quantum mechanics -aiming to describe particles with no spatial extension- one begins with the Hamiltonian of the classical system and proceeds in quantization by promoting the classical momentum and position to non-commuting operators. Aspiring to describe spatially extended but at the same time very small physical objects, or fields[94], physicists considered the Hamiltonians of classical fields and proceeded to what is sometimes referred to as second quantization. Roughly, the process of second quantization involves treating fields as though they were operators and thus giving them the status of quantum fields. So, in the case of fields, we quantize the field and its derivatives rather than the position and the momentum of the particle. Gauge quantum field theories evolved from constrained Hamiltonian systems, something that one familiar with the techniques used in classical quantum mechanics would expect. In this thesis we will not discuss the quantization processes of fields nor the problems that are involved in it[95]. However, we will probe deeply into the Hamiltonian systems, first, and then into their 'heirs', the fibre bundles, that are used in these theories. In this sense, the discussion that follows will be restricted to classical systems, yet we have at the back of our minds the

---

[94] The wavelengths of the objects under consideration are of the order of $10^{-14} - 10^{-16}$ meters.

[95] Dirac, in his *Lectures on Quantum Mechanics*,Belfer Graduate School of Science, Yeshiva University, New York, 1964, writes about these problems: "Some people are so much impressed by the difficulties of passing over from Hamiltonian classical mechanics to quantum mechanics that they think that maybe the whole method of working form Hamiltonian classical theory is a bad method". And further down, commenting on some alternative approaches, he continues: "Still, I feel that these alternative methods, although they go quite a long way towards accounting for experimental results, will not lead to a final solution to the problem. I feel that there will always be something missing from them which we can only get by working from a Hamiltonian, or maybe from some generalization of the concept of a Hamiltonian. So I take the point of view that the Hamiltonian is really very important for quantum theory".

fact that quantization is only a step further and that one way of doing it is by using the so called canonical quantization procedure, which is based on Dirac's treatment of constrained Hamiltonian systems.

In field theory, the typical procedure is the following. We begin with the Lagrangian of our system and not with the Hamiltonian. The reason for this is that if we started with the Hamiltonian it would be difficult to formulate the conditions for the theory to be relativistic[96], so we begin with the Lagrangian, construct an invariant action integral and proceed to get the Hamiltonian and equations of motion for the dynamic variables of the system/structure. One might ask why bother and make the transition from Lagrangian to Hamiltonian at all. After all the Euler-Lagrange equations are equally good. But then, this is just an intermediate step before quantization, and in order to quantize we need quantities that are first order in time derivatives; these quantities we get from the Hamiltonian systems. Thus, the route starts from a Lagrangian and a relativistically invariant action integral, continues through the Hamiltonian formulation and finishes at the quantization of the system. In passing, it is worth mentioning here that the two formulations -the Lagrangian and the Hamiltonian- are mathematically equivalent and the transition from the one to the other is done with the help of the so called Legendre transformations. As we shall see in a while, the invertibility or not of Legendre transformations is related closely to the presence or not of further relations that may hold between the canonical variables of the theory, which in turn determine whether the mathematical description of the physical structure is deterministic or not and allows for the description of interactions.

---

[96] For a more extensive discussion, see Dirac, p.5.

In quantum field theory we deal with systems with infinite degrees of freedom, which could be viewed as a generalization of systems with a finite number of degrees of freedom. N particles or degrees of freedom give a phase space -i.e. space of all possible position and momenta of the N particles- of dimension $2N$, which is a $2N$-dimensional manifold. A field could be considered as the limiting case of an N-particle system as $N \rightarrow \infty$. In this case, the phase space is an infinite dimensional manifold.

A general way to think about a Hamiltonian system is as a triplet (M, ω, H), where (M, ω) is a symplectic manifold -corresponding to the phase space- with a non-degenerate two-form ω and H is a distinguished C^∞ function on (M, ω), which induces a global Hamiltonian vector field X_H on M. The integral curves of the vector field X_H are called the dynamical trajectories of (M, ω, H) and are the solutions to Hamilton's equations. In other words, what this means is the following. Consider that we want to describe a physical system with, say, N degrees of freedom. The whereabouts of such a system will 'define' the so called dynamical trajectories on the $2N - \dim$ manifold M of the phase space of the system. For a Hamiltonian system, this phase space is the cotangent bundle T^*Q of its $N - \dim$ configuration space Q. The dynamical trajectories depend, of course, on the Hamiltonian of the system, which thus defines a vector field, and could be visualized as 'lines' in that $2N - \dim$ cotangent bundle. We use the lower case letter q to denote coordinates and p to denote momenta and their number represents of course the dimension of the phase space as well as the degrees of freedom of the physical system/structure. The Lagrangian of the system, on the other hand, defines a vector field on the tangent bundle TQ, which constitutes the dual space to that of T^*Q and the elements of one space

are mapped onto the other by what we have called the Legendre transformation, and the fact of whether it is or it is not invertible is related to the presence of constraints in the system. When there are *certain* constraints present, not only is the determinant of the transformation zero but also the two-form defined on the manifold[97] is degenerate. In this case, the manifold is said to be a presymplectic manifold.

A classical example of a constrained physical system consists of a bead confined to move round a circular ring which has only one degree of freedom on the configuration space, rather than the original three of the spatial coordinates. This reduction in the original number of the canonical coordinates has the following results.

The accelerations at a given time are not uniquely determined by the positions and the velocities at that time and the general solution of the equations of motion contain, therefore, arbitrary functions of time. The resulting non-uniqueness of the equations of motion entails two things. First, that the state of the system is not uniquely determined by the equations of motion and the initial conditions. For the given system, this means that although we know where on the ring we may find the bead, we could find the whole system at any height from the origin. Second, that the determinant of the Legendre transformation -which is of the form det $\left(\frac{\partial^2 L}{\partial x^a \partial x^b}\right)$- is zero, hence the transformation is non-invertible. The importance of these two outcomes becomes very prominent in field theory, in Hamiltonian systems constrained by gauge symmetries, which we will examine shortly. Before we do that, however, we need to clarify what we mean in general by the term constrained

[97] The two-forms are mathematical objects that are dual to the vectors and while the vector fields correspond to what we would understand as the 'position vectors', the forms -and the connections to which they give rise- inform us about the 'motion' of the objects that are defined on the manifold.

Hamiltonian systems. In the physics literature the notion of constraint is a general one that embraces classical cases like the example we gave above as well as other kinds of constraints, like the ones related to gauge symmetries. According to Henneaux & Teitelboim, "the presence of arbitrary functions of time in the general solution of the equations of motion implies that the canonical variables are not all independent. Rather, there are relations among them called constraints. *Thus, a gauge system is always a constrained Hamiltonian system.* The converse, however, is not true. Not all conceivable constraints of a Hamiltonian system arise from gauge invariance"[98]. However, for some in the philosophical literature[99] "a constrained Hamiltonian system is a gauge theory (N, σ, H) where (N, σ) is a regular submanifold of a symplectic manifold (N, ω)". We favor the former, more general -though less formal- account for what constitutes a constrained Hamiltonian system and we consider a gauge theory to be a field theory whose action is invariant under gauge transformations.

For a system with infinite degrees of freedom and with a gauge symmetry, on the other hand, the constraints express relations between the original infinitely many degrees of freedom that define equivalence classes on the phase space (which we will call gauge orbits). The idea is that within each equivalence class the physical system does not change although the variables associated with it do. The presence of those further relations manifests itself as follows. Given the Lagrangian L describing a physical system, the Hamiltonian H is defined as

$$H = \dot{q}^n p_n - L$$

---

98 Henneaux & Teitelboim, *Quantization of Gauge Systems*, p.4.

99 As in Belot's PhD Thesis, for instance.

where $\dot{q}$ are the velocities of the canonical coordinates while p are the canonical momenta and are defined as

$$p_n = \frac{\partial L}{\partial \dot{q^n}}.$$

If we vary H we get

$$\delta H = \dot{q^n}\ \delta p_n + \delta\ \dot{q^n}\ p_n - \delta\ \dot{q^n}\ \frac{\partial L}{\partial\ \dot{q^n}} - \delta q^n \frac{\partial L}{\partial q^n} = \dot{q^n}\ \delta p_n - \delta q^n \frac{\partial L}{\partial q^n}$$

from which we see that the $\delta\ \dot{q^n}$'s appear only implicitly since $p_n = p_n(q, \dot{q})$. This means that the Hamiltonian is a function of the p's and the q's only and not of the velocities. When the generalized momenta are not all independent functions of the velocities, there are certain relations connecting the momentum variables and are of the type $\varphi_m(p, q) = 0$[100]. One can understand these relations as resulting from the variation of the action and the relation $\frac{d}{dt}(\frac{\partial L}{\partial \dot{q_n}}) = \frac{\partial L}{\partial q_n}$ that follows from a variation of L. When the Lagrangian does not depend explicitly on the coordinate q_n, then $\frac{d}{dt}(\frac{\partial L}{\partial \dot{q_n}}) = 0$ and this results in a relation of the type $\varphi_m(p, q) = 0$. These are what we call first class constraints[101] and according to Noether's theorems they are the reason for conservation of the generalized momenta associated with them. Then, the total Hamiltonian of the system -which is not uniquely determined anyway- is $H_T = H + u^m \varphi_m$. Further, imposing the condition that the equations of motion do not involve inconsistency, from the Poisson bracket of $[H_T, \varphi_m] \approx 0$ we get one out of the three following possibilities: $0 = 0$, which is satisfied identically with the help of primary constraints, or $\chi(q, p) = 0$, or neither. The equations of the form $\chi(q, p) = 0$ imply that

---

[100] This corresponds to the property that the Lagrangian is uncertain to within a total time derivative of an arbitrary function of the coordinates, possibly the momenta, and the time.

[101] Goldstein, in his *Classical Mechanics*, calls them holonomic constraints and their conjugate coordinates cyclic (p.11, 55).

we have further constraints on the Hamiltonian. These are known as secondary constraints and they differ from the primary in that the primary constraints are direct consequences of the definition of momentum, while to derive the secondary, one uses the equations of motion as well. On the other hand, any dynamical variable $R(p,q)$ is said to be first-class if it has zero Poisson brackets with all the primary constraints, i.e. $[R,\varphi_j] \approx 0$[102], $j = 1,...,\mathbb{J}$. Otherwise, R is said to be second class. The constraints that are of interest to us are the *primary first class constraints*, which are arbitrary functions of time, they are the generating functions of what Dirac calls *infinitesimal contact transformations*[103] and fall under the more general heading of symmetry transformations since they lead to changes in p's and q's that do not affect the physical state of the system. The transformations we call gauge are of this type.

One thing we get from the discussion above is that the Hamiltonian $H = \dot{q}^n\, p_n - L$ "is well defined only on the submanifold defined by the primary constraints and can be extended arbitrarily off that manifold. It follows that the formalism should remain unchanged by the replacement

$$H \rightarrow H + u^m(p,q)\varphi_m\text{"}^{104}.$$

With the addition of the new variables $u^m(p,q)$ we restore invertibility of the Legendre transformation but the cost we actually pay is that there are now many sets of values of the canonical variables that represent a given physical state. This means that if we were

---

[102] The symbol '$\approx$' reads 'weakly equal' and it means that one has to work the Poisson bracket first and then take the constraints to be equal to zero; in other words, one considers the Poisson brackets on the constraint surfaces.

[103] These are what Goldstein calls canonical transformations and points out the fact that the terminology in the literature is not standard (*Classical Mechanics*, p.381).

[104] Henneaux & Teitelboim, *Quantization of Gauge Systems*, p.11.

given an initial set of values for our canonical variables at some time t_1, we would not be able to determine *uniquely* the physical state of the system at other times. This kind of indeterminism is inherent to the formulation of the theory and for that reason different from indeterminism that results from the random nature of certain physical phenomena, like radioactivity for example, or probabilism, as this manifests in quantum mechanics, say. One last consequence of the non-invertibility of Legendre transformations is that the Lagrange equations of motion are non-integrable. All these consequences, along with attempts to cure the lack of indeterminism will be discussed in the rest of this chapter and in the next.

For the transition from a system/structure with finite -say n- to a system/structure with infinite degrees of freedom, we take the limit $n \rightarrow \infty$ and $\varphi(x^\mu)$, $\partial_\mu \varphi$, instead of p's and q's, where x^μs play the role of parameters -a role similar to that of t in the finite case. So far as the constraints are concerned, in the infinite case they take the form of divergence conditions[105].One important thing to bear in mind is that , as Dirac points out,"from a practical point of view, one can tell from the general transformation properties of the action integral what arbitrary functions of the time will occur in the general solution of the equations of motion. To each of these functions of the time there must correspond some primary first class constraint"[106]. To illustrate what we have just said, we proceed now to consider an example of an infinite dimensional Hamiltonian system with constraints, namely, the classical free electromagnetic field, which is of major interest to us for reasons that will become clear later.

---

105 See, for example, Goldstein, pp.555-6.

106 Dirac, *Lectures on Quantum Mechanics*, p.19.

3.2.1 The Free Electromagnetic Field

The dynamical coordinates in this case are the potentials $A_\mu(x)$, where we will consider x to stand for the three spatial coordinates x^1, x^2, x^3, at a given time $x^0 = t$. The generalized velocities are, then, the time derivatives $\partial_0 A_\mu(x)$ of the dynamical/generalized coordinates. The Lagrangian density is given by $\mathcal{L} = -\left(\frac{1}{4}\right) F_{\mu\nu}F^{\mu\nu}$, where $F_{\mu\nu} = A_{\mu,\nu} - A_{\nu,\mu}$ is what we call the electromagnetic field tensor. The Lagrangian of the system is $L = \int \mathcal{L} d^3x = -\left(\frac{1}{4}\right) \int F_{\mu\nu}F^{\mu\nu} d^3x$ and as we can see it does not depend explicitly on the generalized coordinates. Hence, we are expecting that certain constraints will apply. If we take the variation of the Lagrangian, now, and we define the momenta B^μ as $B^\mu = F^{\mu 0}$, we can see that from the antisymmetry of the electromagnetic field tensor follows immediately that $B^0(x) = 0$. This is a primary constraint for which we can write $B^0(x) \approx 0$ and given that x represents a point in a three-dimensional Euclidean manifold, the relation refers to an infinity of primary constraints: each value of x will give a different primary constraint!

The other momenta, $B^r(x) = F^{r0} = \partial^r A^0 - \partial^0 A^r$, $r = 1, 2, 3$, are just the components of the electric field and if we rewrite the Lagrangian applying the constraint, we may get an expression for the Hamiltonian that does not involve velocities any more, just the rest of the generalized momenta -i.e. spatial derivatives of the field[107]. As it turns out, the variables A_0, B_0 are not of any physical significance and, therefore, they are redundant[108]. This redundancy is precisely the result of the constraints that apply in the system and it

[107] The Hamiltonian we get using the relation $H = p\,\dot{q} - L$ and in this case $H = -L$ because it does not depend explicitely on the generalized coordinates.

[108] The electromagnetic field has only two (transverse) components, as revealed by the two directions of polarization of light.

is related, as we shall see in the following section, with certain symmetries and symmetry transformations, known as gauge, that leave the action of the system invariant.

From a matching relativistic treatment of the same system we get the following results. The relativistically invariant Lagrangian is

$$L = \int \mathcal{L} d^4x = -\left(\frac{1}{4}\right) \int F_{\mu\nu} F^{\mu\nu} d^4x$$

and apparently it is invariant under the transformation

$$A_\mu \rightarrow A' = A_\mu + \Lambda_\mu.$$

which we call global gauge transformation[109]. The very fact of invariance of the Lagrangian under the above symmetry transformation entails that the system is constrained.

3.3 Symmetries, Conserved Quantities and Interactions

The notion of symmetry is very important in contemporary physics for two reasons. One reason, which despite its importance has been rather neglected in the philosophical literature, is that Noether's theorems associate symmetries with conserved quantities and conservation principles. The second is that the so called local symmetries allow for coupling terms that are interpreted as interactions. But let us examine each of these two reasons in some depth.

---

[109] This type of transformation is called global because the parameter Λ_μ of the transformation does not have any spacetime dependence.

3.3.1 Noether's First Theorem and Conservation Laws

As we have already mentioned, a Hamiltonian system with primary first class constraints is subject to gauge transformations that leave the physical state of the system unaffected. Noether's three theorems[110] connect symmetries of Lagrangian systems[111] with conservation laws as follows. The first theorem concerns systems with continuous symmetries depending on constant parameters and it states that in such a system, and given that all (matter) fields that are affected by symmetry transformation satisfy the Euler-Lagrange equations, we can derive a continuity equation. From this equation we can get a conservation law by performing an integration. Examples of such conservation laws are those of energy, momentum and electric charge. From an algebraic point of view, the terms that appear in the conserved currents or in the continuity equations are the generators of the infinitesimal symmetry transformations that leave the physical system unaffected[112]. Taking as an example a physical system/structure involving complex scalar fields we will be able to see how symmetries of the mathematical structure deliver conservation laws for energy-momentum and something that we would like to identify with the electric charge.

Consider a scalar field of the form[113]

$$\varphi = \left(\frac{1}{\sqrt{2}}\right)(\varphi_1 + i\varphi_2)$$

---

[110] As a matter of fact, it is only the first two theorems that were derived by Noether herself, the derivation of the third was due to Utiyama. Nevertheless, all the three of them follow from Noether's variational problem. For an extended discussion see Brading and Brown (2001).

[111] Note that in order to discuss Noether's theorems we go back to the Lagrangian systems. This is not a drawback, as it may seem at the beginning, since the two approaches are in fact equivalent. It is only a matter of convenience which one might choose.

[112] These infinitesimal transformations can of course be integrated to give us the finite symmetry transformations.

[113] In this presentation, we follow Ryder's *Quantum Field Theory* pp.93.

$$\varphi^* = \left(\frac{1}{\sqrt{2}}\right)(\varphi_1 - i\varphi_2)$$

where $\varphi = \varphi(x)$ and $\varphi^* = \varphi^*(x)$ we regard as independent fields and 'trace out' a region R of the $4-$ dim spacetime manifold. Then a relativistic invariant Lagrangian density that we could write for this field is the following:

$$\mathcal{L} = (\partial_\mu \varphi)(\partial^\mu \varphi^*) - m^2 \varphi^* \varphi$$

and the Euler-Lagrange equations of motion, which are derived by requiring $\delta S = \delta \int \mathcal{L} d^4x$, give the two Klein-Gordon equations:

$$\begin{aligned}(\Box + m^2)\varphi &= 0 \\ (\Box + m^2)\varphi^* &= 0.\end{aligned}$$

This is done as follows. Varying the action integral with respect to both the coordinates and the field -a variation which vanishes at the boundary ∂R of the region R- we get:

$$\delta S = \int_R \{\frac{\partial \mathcal{L}}{\partial \varphi} - \partial_\mu [\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}]\delta\varphi d^4x\} + \int_{\partial R} [\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}\delta\varphi + \mathcal{L}\delta x^\mu] d\sigma_\mu + \text{complex } conjugate =$$

$$\int_R \{\frac{\partial \mathcal{L}}{\partial \varphi} - \partial_\mu [\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}]\delta\varphi d^4x\} + \int_{\partial R} \{\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}[\delta\varphi + (\partial_\nu \varphi)\delta x^\nu] - [\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}\partial_\nu \varphi - \delta^\mu_\nu \mathcal{L}]\delta x^\nu\} d\sigma_\mu + c.\ c.$$

The boundary term vanishes anyway because $\delta\varphi = 0$ and $\delta x^\mu = 0$ there. So, from the requirement that the action is stationary we get the Euler-Lagrange equations of motion for the two fields, while for the boundary term we can write the following equation:

$$\int_{\partial R} \{\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}[\delta\varphi + (\partial_\nu \varphi)\delta x^\nu] - [\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi)}\partial_\nu \varphi - \delta^\mu_\nu \mathcal{L}]\delta x^\nu\} d\sigma_\mu + c.\ c. = 0.$$

Taking the total variation of the field φ to be $\delta\varphi + (\partial_\nu\varphi)\delta x^\nu = \Delta\varphi = \Phi_\mu\delta\omega^\mu$, where $\delta\omega^\mu$ is an arbitrary constant variable, and $\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi)}\partial_\nu\varphi - \delta^\mu_\nu\mathcal{L} = \vartheta^\mu_\nu$, the equation above becomes

$$\int_{\partial R}\{\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi)}\Delta\varphi - \vartheta^\mu_\nu\delta x^\nu\}d\sigma_\mu + c.\,c. = 0$$

Suppose, now, that the transformations under which the action integral is invariant take the form

$$\Delta x^\mu = X^\mu_\nu\delta\omega^\nu \text{ and } \Delta\varphi = \Phi_\mu\delta\omega^\mu$$

Then

$$\int_{\partial R}\{\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi)}\Phi_\nu - \vartheta^\mu_\kappa X^\kappa_\nu\}\delta\omega^\nu d\sigma_\mu + c.\,c. = 0$$

which, because the parameter of the transformation $\delta\omega^\nu$ is arbitrary, we can rewrite as

$$\int_{\partial R}$$ [illegible] $$\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi)}$$ [illegible] (*)

As we can see, the J^μ_ν contains a term emerging as a result of spatiotemporal variation and a term coming forward as a result of variation of the φ-fields. Applying Gauss's theorem we finally get

$$\int_{\partial R} J^\mu_\nu d\sigma_\mu = \int_R \partial_\mu J^\mu_\nu = 0$$

from which follows that $\partial_\mu J^\mu_\nu = 0$ since R is arbitrary. This last equation tells us that we have a conserved current J^μ_ν which is the result of the invariance of the action under the transformations $\Delta x^\mu = X^\mu_\nu\delta\omega^\nu$ and $\Delta\varphi = \Phi_\mu\delta\omega^\mu$. If we integrate this current over a spacelike hypersurface σ_μ we get a conserved quantity, or charge,

$$Q_\nu = \int_\sigma J^\mu_\nu d\sigma_\mu$$

---

[114] We also get a similar result involving the complex conjugate field φ^*.

as expected from Noether's first theorem. The relation $\partial_\mu J^\mu_\nu = 0$, which is a divergency term, apparently represents a constraint of our system and to classify it one has just to check its commutation relation with the Hamiltonian of the system, but this is beyond the scope of this presentation. Now, the question is what does this conserved quantity represents, or to put it in the terminology of the second chapter, is the relation $\partial_\mu J^\mu_\nu = 0$ mapped onto some physical relation or does it belong to the surplus structure? The transformation of the coordinates, when interpreted in an active way, corresponds to a change of the spacetime region on which our physical structure is defined. Consider now that the transformation of the coordinates is such an active translation, while for the φ field $\Delta\varphi = 0 \rightarrow \Phi_\mu = 0$. Then, we can recognize the energy-momentum tensor in the generator X^μ_ν of the infinitesimal transformation $\Delta x^\mu = X^\mu_\nu \delta\omega^\nu$. Hence, the conserved current in this case is nothing other than the energy and the linear and angular momentum of the system.

Consider now that $\Delta x^\mu = 0$, i.e. that $X^\mu_\nu = 0$, and that the φ fields undergo the transformation $\varphi \rightarrow e^{-i\Lambda}\varphi$ and $\varphi^* \rightarrow e^{i\Lambda}\varphi^*$. The infinitesimal form of this transformation is

$$\delta\varphi = -i\Lambda\varphi \text{ and } \delta\varphi^* = i\Lambda\varphi^*$$

so that

$$\Phi = -i\varphi \text{ and } \Phi^* = i\varphi^*.$$

Using the general relation for J^μ_ν that we derived before (equation $(*)$) we get

$$J^\mu = -i\varphi\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi)} + i\varphi^*\frac{\partial\mathcal{L}}{\partial(\partial_\mu\varphi^*)}.$$

This relation, in conjunction with the Klein-Gordon field equations, gives us $\partial_\mu J^\mu = 0$ and a corresponding conserved quantity

$$Q = i \int (\varphi^* \frac{\partial \varphi}{\partial t} + \varphi \frac{\partial \varphi^*}{\partial t}) dV.$$

This conserved quantity, which we would like to identify with electric charge[115], appears as a result of a symmetry transformation, a gauge transformation with constant transformation variable, which represents a rotation in an internal space. This internal rotation does not seem to correspond to anything physical, and so it is responsible for ambiguity of representation of the third kind. Still these internal symmetry transformations will play a crucial role in the description of interactions when we allow the parameter of the transformation to vary with spacetime, as we shall see shortly. As a last remark let us mention, again, that the relation $\partial_\mu J^\mu = 0$ is a constraint whose nature we could identify by checking its Poisson brackets with the Hamiltonian of the system.

3.3.2 Noether's Second and Third Theorems and Interactions

The second and the third theorems concern the case of symmetry transformations whose parameters depend smoothly on arbitrary functions of spacetime and their derivatives. The general expression we get from the variational problem in a case like this consists of an interior contribution and of a boundary contribution that must vanish independently. When we require each of them to vanish, we get Noether's second theorem from the vanishing interior contribution and the third theorem from the vanishing boundary contribution[116]. Brown

[115] This quantity, as a matter of fact, does not contain anything that could be identified as the charge of the field φ, nor anything that could be interpeted as quantization of the charge. In the following section, when we talk about 'local' gauge transformations we will get back to this point.

[116] For a detailed derivation see Brading and Brown, *Noether Theorems and Gauge Symmetries*.

and Brading have shown in their 2001 that from the third theorem follow three equations which could be interpreted as follows. The first one says that given the gauge field equations, a conserved current expressed in terms of the matter fields may be derived, which is independent of the matter field equations. The second says that given the gauge field equations, this conserved current acts as the source of the gauge fields. Finally, the third expresses a constraint on the form of the gauge fields. The second theorem combined, with the first of the three equations of the third theorem shows that, given the matter field equations ,another conserved current may be derived independently of the gauge field equations So, aside from the conservation relations, in the case of local gauge transformations[117] we also get coupling terms, that is terms that join gauge with matter fields and it is precisely these terms that can be interpreted as describing interactions. To illustrate all these, we will cite as example the case of the complex scalar field and the electromagnetic field in one system.

For a complex scalar field in an electromagnetic field we could begin with a Lagrangian density that combines $\mathcal{L} = -(\frac{1}{4})F_{\mu\nu}F^{\mu\nu}$, the Lagrangian density of the free electromagnetic field as we have seen, with the one of the free scalar field, namely $\mathcal{L} = (\partial_\mu\varphi)(\partial^\mu\varphi^*) - m^2\varphi^*\varphi$. So, the Lagrangian density $\mathcal{L}$ of the system takes the form

$$\mathcal{L} = (\partial_\mu\varphi)(\partial^\mu\varphi^*) - m^2\varphi^*\varphi - \left(\frac{1}{4}\right)F_{\mu\nu}F^{\mu\nu}.$$

[117] This kind of gauge transformations are called local because the parameter(s) of the transformation have spacetime dependence and not because they are related to localized currents or local conservations laws. There are two different issues here, as a matter of fact, to which will come back in the next chapter.

Apparently, this system is invariant under global transformations of both the scalar and the gauge fields, but if we consider the following local transformations,

$$\varphi \rightarrow e^{-i\Lambda(\mathbf{x},t)}\varphi$$

$$\varphi^* \rightarrow e^{i\Lambda(\mathbf{x},t)}\varphi^*$$

$$A_\mu \rightarrow A_\mu + \partial_\mu \Lambda(\mathbf{x},t)$$

this is not so. Although each of the two constituent-Lagrangians are invariant under both global and local transformation, the derivatives of the scalar fields in total Lagrangian 'hit' the transformation parameter and produce extra terms. But this downside can be sorted out if we make amendments to our original Lagrangian. And to do this, we only need to replace the partial derivatives ∂_μ by what we call the covariant derivatives D_μ which are of the form

$$D_\mu = \partial_\mu - iA_\mu.$$

The presence of this extra term restores invariance in the Lagrangian density, which now takes the form

$$\mathcal{L} = (D_\mu\varphi)(D^\mu\varphi^*) - m^2\varphi^*\varphi - \left(\frac{1}{4}\right)F_{\mu\nu}F^{\mu\nu}.$$

From Noether's second and third theorems and the Lagrangian density above we get the following result

$$\partial_\mu j^\mu = \partial_\mu\{\frac{\partial L}{\partial A_\mu} - \partial_\nu \frac{\partial L}{\partial(\partial_\nu A_\mu)}\} = 0$$

when the matter field Euler-Lagrange equations hold. But also we can arrive at the conserved current when the gauge field Euler-Lagrange equations hold. Hence, we can conclude that although the Euler-lagrange equations of the matter fields are sufficient for the derivation of a conserved current, they are not necessary. This divergency condition rep-

resents a constraint, which we were able to derive as a consequence of the symmetries of the Lagrangian and because we used the Euler-Lagrange equations in its derivation, it is a secondary one. Moreover, since

$$j^{\mu} \stackrel{\circ}{=} -\frac{\partial L}{\partial A_{\mu}} = \partial_{\nu} F^{\mu\nu},$$

where the symbol '$\stackrel{\circ}{=}$' means that the equality holds independently of any Euler-Lagrange equations,we can identify this conserved current with the electric current, since what we can read off from this equation is that the conserved current is the source of the electromagnetic field. This last result, as Brading and Brown point out, is an instance of a more general result that follows from Noether's third theorem given satisfaction of the Euler-Lagrange equations of all those fields whose transformations depend on the derivatives of the arbitrary variables-functions, i.e. on $\partial_{\mu}\Lambda(\mathbf{x}, t)$. This result gives us what they call *coupled field equations* which we then interpret as interaction terms. Hence, that's how interactions arise as a result of the local gauge invariance of the system.

One thing worth noticing here is that, as a matter of fact, the electric charge or coupling constant q does not come up as a consequence of gauging. The only reason why it should appear is because we want the conserved quantity that we calculate from the conserved current -by integration- to represent the total charge of our system. Hence the coupling constant is introduced in the mathematical structure as a further constraint imposed by 'external' physical requirements.

Let us conclude this section by connecting it also to the discussion of the previous chapter. The physical system we want to describe, here, consists of matter-fields that interact electromagnetically, while the mathematical structure we are using is this of the con-

strained Hamiltonian systems. The concrete mathematical structure we employ here is an infinite dimensional manifold, a presymplectic manifold to be precise, and what happens is that we map a state of the physical system to a point in the manifold, which is a concrete mathematical object. The presence of constraints in the mathematical algebraic structure means that we have a plethora of mathematical objects in the manifold that constitute an equivalence class onto which a single state of a physical object is mapped. This, of course, is an instance of the third type of ambiguity we have mentioned, which here we call symmetry because the changes it dictates do not affect the physical system we are studying. This ambiguity is also related to the notion of surplus structure in the sense that the Hamiltonian systems that we choose each time to represent a physical system have more degrees of freedom than the ones required by the physical system for its description. This is reflected by the presence of redundant degrees of freedom, which one could claim belong to the so-called surplus structure.

Yet, it is precisely this ambiguity, the presence of symmetry, that delivers conserved currents and coupling terms in the algebraic structure. We use these conserved quantities, along with a further, external, requirement to represent sources of the interaction-fields, while the coupled terms that arise when we require invariance under the symmetry transformations represent interactions.

3.3.3 Symmetry, Ambiguity of Representation and Indeterminism

The very fact that in constrained Hamiltonian systems we have more field-degrees of freedom than the ones we need in order to describe the physical system entails interaction

terms, as we have seen. On the other hand, though, it conceals a lack of determinism which is considered by many to be a problem. Let us see how this indeterminism comes about, first, and then discuss possible attitudes towards it.

The issue is that since we have at our disposal more coordinates than we need, the structure is inevitably non-deterministic. The initial value problem is underdetermined and hence the time evolution of the physical system is not uniquely determined. One way to understand this is by considering that for each symmetry of the mathematical structure, there are certain equivalence classes defined in it. These classes in the case of gauge symmetries are also called gauge orbits. Now, the idea is that all elements of a class correspond to the same state of the physical system they represent, hence if we know where we started, we can never be sure on which element of a class the time evolution of the system will take us to. For structures with gauge symmetries, a remedy would be to fix the gauge. The gauge fixing is basically to choose one out of the infinitely many gauges of an equivalence class and treat the evolution of the physical structure taking it as constant. This solution, however tempting, involves a problem that will become clearer in what follows, after we have talked about fibre bundles. For the time being, though, suffice it to say that in some cases we are not able to specify the gauge throughout the spacetime manifold, so we cannot fix the gauge uniquely.

Another way to treat indeterminism is by considering that the actual physical objects are described by gauge invariant quantities. This, however, deprives our explanations from causal pictures and, as we shall see in the next chapter, leads to non-locality. But for now let us just say that in this case the problem is that, apparently, there is more information

in the structure-as-a-whole than the nearby neighboring points can give us which results in the problem of non-locality.

3.4 Local Symmetries Giving Rise to Interactions

In the discussion above we pointed out that constrained Hamiltonian systems are associated with symmetries that may be of a local or of a global nature and with conserved currents and quantities. We also saw that it is global symmetries that generate currents and local symmetries that produce coupling or interaction terms, although local symmetries are also associated with conserved currents but for currents to emerge out of the variations we need to take a few more steps. To our knowledge, the use of the terms 'local' and 'global' has created some sort of confusion in the literature which we would like to clarify and which resulted in a misunderstanding that we will try to put an end to before we proceed any further. The main culprit for this confusion and misunderstanding is that while there are two different notions of locality that arise in the discussion of symmetries and interactions, they are often muddled. So, what is the difference between local and global in this context and why local symmetries as opposed to global? The answer to the question 'why local' comes in two parts. The first part is concerned with the notion of charge and its local conservation, while the second is concerned with the notion of interaction. So, at this point we should distinguish between the two 'localities' that have appeared so far, so that the differences and the relations between '*local charge conservation*' and '*local symmetries*' become clear.

'Local charge conservation' refers to conservation of charge, as the words suggest, which is described using currents localized in spacetime. The point why we should expect the charge to be conserved locally may be argued for using relativistic considerations, and this is typically done as follows[118]. Special relativity theory tells us that it is impossible to tell the difference in physical laws whether we are moving or not. If conservation of electric charge was non-local, that is if charge was to disappear from one place and simultaneously appeared in another, this would be so for just one special observer. For any other observer in relative motion to the special one, appearance and disappearance would not be simultaneous. Therefore, one could tell by this difference whether the two observers were in relative motion to each other or not. But according to relativity theory it is impossible to tell, therefore the special observer cannot exist and hence the conservation of electric charge must be local.

Meanwhile, Noether's Theorems tell us that local conservation laws arise as a result of symmetries which may be global as well as local -in the latter case, Noether's 2nd theorem gives a generic relation-constraint which is usually read as a linear combination of identities and conservation laws[119]. So, if we describe the events using the notion of symmetry, we get conservation laws that allow for local conservation of the electric charge, that is to say, we get currents which describe how charge is transported from one 'place and time' to another continuously. Taking global symmetries into account, the conservation currents and the conserved quantities follow as a result of Noether's 1st theorem. From this

---

[118] For further details see Aitcison & Hey, *Gauge Theories in Particle Physics*, or Feynman, *The Character of Physical Law.*

[119] For a detailed discussion see M. Bremer, *Notes on D=11 Supergravity* and C. Brading, *Which Symmetry?*

perspective, not only the total charge is conserved but also what the charge does complies with relativity principles and it satisfies relativistic equations, which is what one would expect it to do. So, through Noether's theorems, local conservation laws are derived, as we have seen: in the case of global gauge transformations Noether's first theorem guarantees that there will be some conserved currents that satisfy relativistic requirements and are local in this sense, while in the case of local gauge transformations, her second theorem discloses some identities through which we may identify conserved quantities which are also local in the same sense. Global or local symmetries, therefore and Noether's theorems are sufficient for derivation of localized conservation currents and conserved quantities. But there is a difference between global and local symmetry transformations as to what kind of physical structures they may describe. What we need to clarify next is precisely the meaning of and the differences between the notions of global and local symmetry transformations.

When we talk about 'local symmetry transformations' we actually refer to transformations of the Lagrangian and the equations of motion of our system with a transformation parameter that has spacetime dependence and thus may vary as we move from spacetime point to spacetime point, hence they are local in this sense. On the other hand, the parameter in the so called global symmetry transformations has no spacetime dependence and therefore once it is chosen it is fixed throughout the spacetime manifold. The transformations we have in mind here take place in some internal space, not in the actual spacetime, and they are not directly related to local charge conservations. So, arguments that try to employ local charge conservation as a justification for the use of local symmetry transformations just mix up two different things that are not relevant to each other in the sense

sometimes claimed. Both local symmetries as well as the global ones account for continuity of 'charge transportation'. Nevertheless, considerations of global symmetries are unable to account for any interactions and hence it is only local symmetries that give rise to coupling terms that are mapped to interactions.

We would like to emphasize once again that the presence of interaction terms is necessary, since it is through physical interactions that we observe the physical entities. Within the context of the Hamiltonian formalism, interaction terms appear straightforwardly when we require certain global symmetries of the theory to acquire a local character. So, given the mathematical tools we currently have, we may describe interactions if we use local symmetries. The use of local gauge symmetries is a sufficient and consistent way of 'generating' interaction terms and, therefore, of describing/representing interactions.

This doesn't mean to say that the action of gauge fields -as thus dictated by the theory- is local. As we shall see in the next chapter, it is not possible to give an interpretation that allows for local action of the gauge fields and this results from the fact that Legendre transformations are non-invertible and therefore the equations of motion are non-integrable. However, this is, once again, a different issue that does not interfere with their local character as we have expressed it here.

3.4.1 Spacetime, Matter, Interactions and Numbers

In (quantum) field theory, the objects or fields which eventually may be interpreted as elementary particles and carriers of the forces, are rather elaborate objects with various different properties that need to be taken into account. All these properties indicate how

they interact with each other and, as a consequence, they must manifest themselves in their mathematical description. Revealing just the spatiotemporal whereabouts of physical objects is not all the information we need nor all we can get. We are arguing, therefore, that the spacetime indices do not give a sufficient description of the fields because other specifications are needed as well. The specifications we are referring to concern physical quantities like spin, weak isospin, strangeness, lepton number, color etc. These other properties which need to be taken into account are successfully described by complicated 'multiple vectors' with both spacetime and tensorial indices. Therefore, interacting fields need both spacetime and further specifications.

Here, we cannot say that we actually *need* tensorial indices because there are other theories, like the ϕ^4 theory, which describe interactions without using them. But the truth is that these other theories make use of mathematical apparatus that is by no means simpler than the tensors, nor more fruitful. In ϕ^4 theory, for example, physicists use Grassmann algebras and some other mathematical artifacts called Grassmann variables in order to describe fermions. These mathematical objects are not easier to handle than tensors and on the top of that they do not have other virtues of tensor calculus. For example, one cannot read directly from a ϕ^4 the difference between matter and interacting fields, nor one can get a unified picture -no matter how inadequate. Hence, this stuff, we argue, is better -although not uniquely- described using differential geometry. The word 'better' in this context means, basically, more convenient from a mathematical point of view as it has a unifying effect and more economical, from a physical point of view, because all the relevant properties are accounted for, interactions arise predictably from the formalism, and the

physically apparent difference between matter and interaction fields[120] is innate in the formalism. Moreover, the heuristic virtues of this formalism have proved unparalleled in both physical and mathematical directions. Towards the mathematical direction, basically all attempts for unification of the fundamental forces -including string theory- have departed from this starting point. And in the physical direction, the experimental verification of the existence of, say, the weak bosons relied heavily on theoretical predictions of the standard model, which is plausibly incorporated in and enriched by the fibre bundle formalism.

One more advantage of the description of a physical structure of interactive fields using differential geometry and fibre bundles as opposed to constrained Hamiltonian systems is that in the first case we have a top-bottom approach, while in the second we have a bottom-up. Let us explain the latter here and leave the former until after we have introduced the fibre bundle formalism. In physics textbooks, usually, they start with the equations of motion of the fields they intend to describe and from them, they build the Lagrangian of the system, from which the equations are derived using variations. If one knows that the physical structures obey certain conservation laws, one makes implicit use of Noether's theorems and searches for the symmetries that are associated with the system. Then, rather than identifying the constraints and hence the symmetries of the system, they first recognize the symmetries and then derive the constraints, mainly in the form of divergency -or conservation- relations. In the case of electromagnetism, at least, they first work out the global symmetry transformations and then impose the requirement that the parameters have spacetime dependence, hence deriving coupling terms to account for interactions. Interac-

[120] Matter fields have mass and are directly observable, while interaction field are usually massless -the weak field aside- and observable through currents.

tion terms are essential because it is only the very presence of the interaction terms which allows us to calculate quantities that are experimentally observable and observed. Let us describe now how electromagnetic interaction terms arise as a result of rendering the gauge symmetry of the classical theory local.

Complex Scalar and Electromagnetic Fields

This is just a simple example of a field with zero spin, which we are using here to illustrate how by making use of the variational principles and the requirement of gauge invariance we may describe interactions. For that reason, we do not deal with global transformations at all; instead, we examine directly the 'local' case[121].

If the scalar field has two components, we may express it as follows.

$$\begin{aligned} \phi &= \tfrac{1}{\sqrt{2}}(\phi_1 + i\phi_2) \\ \phi^* &= \tfrac{1}{\sqrt{2}}(\phi_1 - i\phi_2)\,. \end{aligned}$$

We start off with the simplest action S we can think of, which will give the two Klein-Gordon equations for the ϕ and its conjugate ϕ^*. So, from the Lagrangian density

$$\mathcal{L} = (\partial_\mu \phi)(\partial^\mu \phi^*) - m^2 \phi \phi^*$$

we get the equations of motion

$$\begin{aligned} (\Box + m^2)\phi &= 0 \\ (\Box + m^2)\phi^* &= 0. \end{aligned}$$

[121] For further reading on variational principle and its applications to field theory, see for example Goldstein, *Classical Mechanics,* Guillemin & Sternberg, *Symplectic Techniques in Physics*, Ryder, *Quantum Field Theory* and L.I. Schiff, *Quantum Mechanics.*

We now require from the action S to be invariant under what we call a local gauge transformation with parameter Λ, under which the fields are transformed as follows:

$$\phi \rightarrow e^{-i\Lambda(x^\mu)}\phi \; and \; \phi^* \rightarrow e^{i\Lambda(x^\mu)}\phi^*$$

The infinitesimal form of this transformation is this

$$\delta\phi = -i\Lambda(x^\mu)\phi \; and \; \delta\phi^* = i\Lambda(x^\mu)\phi.$$

The action is no longer invariant under this transformation and this comes as a result of the dependence of Λ on x^μ. As a matter of fact, the change in Lagrangian is

$$\begin{aligned} \delta\mathcal{L} &= \frac{\partial\mathcal{L}}{\partial\phi}\delta\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\delta(\partial_\mu\phi) + (\phi \rightarrow \phi^*) \\ &= \ldots \\ &= i\partial_\mu\Lambda(\phi^*\partial^\mu\phi - \phi\partial^\mu\phi^*) \\ &= J^\mu\partial_\mu\Lambda \end{aligned}$$

To make the action invariant under these transformations, we introduce a new 4-vector A_μ which couples directly to the current J^μ giving an *extra term* in the Lagrangian:

$$\mathcal{L}_1 = -eJ^\mu A_\mu$$

The coupling constant e has units such that A_μ has the same units as $\partial/\partial x^\mu$. For this new field we require that it transforms as follows:

$$A_\mu \rightarrow A_\mu + \tfrac{1}{e}\partial_\mu\Lambda$$

so that

$$\delta\mathcal{L}_1 = -e\left(\delta J^\mu\right)A_\mu - J^\mu\partial_\mu\Lambda$$

But then, in order to counteract the consequences of the transformation on $\mathcal{L}_1$, we add another term to our Lagrangian, namely

$$\mathcal{L}_2 = e^2 A_\mu A^\mu \phi^* \phi$$

for which

$$\delta\mathcal{L}_2 = 2eA^\mu(\partial_\mu \Lambda)\phi^* \phi$$

and hence

$$\delta\mathcal{L} + \delta\mathcal{L}_1 + \delta\mathcal{L}_2 = 0.$$

For the total Lagrangian $\delta\mathcal{L} + \delta\mathcal{L}_1 + \delta\mathcal{L}_2 = 0$ by virtue of our having introduced a field A^μwhich couples to the current J^μ of the complex field ϕ. This Lagrangian is a good candidate for describing interactions - the coupling term $\mathcal{L}_1 = -eJ^\mu A_\mu$ could be interpreted as an interaction term between the current of a field/particle ϕ and the field A^μ which we may manage to interpret as a force field. Actually, it is not difficult to interpret A^μ as the electromagnetic field; one only needs to introduce one more term in the total Lagrangian such that it is gauge invariant and it gives the equations of motion of the electromagnetic field. This term is

$$\mathcal{L}_3 = -\tfrac{1}{4} F^{\mu\nu} F_{\mu\nu}$$

where

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$$

is the electromagnetic field tensor.

From what we have done so far, it comes to light how the electromagnetic field appears as an interactive field by simply demanding invariance of the action under local gauge transformations. The question that may arise here is what is it that it makes it worthwhile

to require local gauge invariance? We would like to be able to answer by saying "a necessity of nature dictated by the gauge principle", but we do not believe that we can, nor can anyone else for that matter. The only argument that comes anywhere near necessity is that all fundamental interactions known to us so far are interactions described in this way. But aside from that, to our view, there are two features that make the requirement of local gauge invariance plausible, although the effortlessness with which the interaction terms appear is not dictated by any physical necessity. First of all, it is the fact that the equations of motion for both the matter and the electromagnetic fields, as well as the interaction terms, arise from the same variational treatment of a single Lagrangian, which is invariant under both spatiotemporal and internal symmetry transformations. In so far as we have accepted that matter fields may be described using variational principles, it is credible to make use of the same technique in order to describe the electromagnetic field and its interactions despite the fact that the two types of field have different properties. The interaction fields behave differently from the matter fields in that the former display a bosonic behavior (associated with integer spin) while the latter a fermionic one (which means half integer spin) and also in that the former often are massless while the latter are usually massive[122]. What is worthwhile, then, in this approach is the fact that by using just one principle - $\delta S = 0$- and the appropriate Lagrangian, one may derive all the equations needed in order to describe a specific kind of physical interactive structures, which takes into account

[122] As a matter of fact, the weak interaction carriers are gauge bosons with mass -the fact that they must have mass is dictated by their short range. In the formalism, the acquisition of mass of the weak gauge bosons is accommodated by what is known as spontaneous symmetry breaking. When the original gauge symmetry is broken, or hidden, the bosons obtain mass; the price to be paid, though, is that another field -the Higgs- appears in the formalism and it requires some counterpart in nature. So far, the existence of the Higgs field has not been confirmed by experiment.

the different properties the two display. So we have a unified treatment of equations of motion and of interactions, which we may say is 'natural' from a mathematician's point of view. That is to say, we derive everything we need deductively, using first principles and 'plausible guesses' with only requirement that must later be justified experimentally.

Furthermore, the internal symmetry which was used in order to derive the interaction terms, that is to say the gauge symmetry, has been known to be an inherent property of the electromagnetic theory since the times of Maxwell. Of course, the A_μ field appears explicitly in this description and the controversy is about whether A_μ itself is a natural field at all. How could we possibly claim that a quantity which is not even gauge invariant is something more than a mathematical artifact? Or, to use our terminology, could we hint, or even more, would it be possible to show that the space where the gauge fields live and are transformed is something more that just the surplus structure, an already elaborate mathematical structure?

It is essential to figure out if the appearance of the newcomer A_μ makes sense in the physics we already have, but for the time being we would like to postpone any arguments about the possible interpretations that one could ascribe to/associate with this (originally mathematical) object. The reason is that we would like first to examine what this field does when we adopt the fibre bundle approach and then try to convince you that what we actually gain is a lot more than what we seem to lose. We will try to argue, then, that the losses are not real losses. What really goes on, as a matter of fact, is that we are just moving away from an old approach giving up some of its limitations -and/or constraints- while at the same time we are embracing a new approach which is much more fruitful in terms of

predictions and explanations, more comprehensive and more open to new perspectives and possibilities.

After having completed this task, we will come back to the issue of whether A_μ is a natural field and then we will consider some possible interpretations of this field and of some other objects that we will have encountered by then. But until then, let us continue our examination of the mathematical properties and relations of these fields.

So far, we have argued that interactions are described successfully and sufficiently[123] using local symmetries. There, the matter fields are described by tensors, while the carriers of the interactions appear as correcting terms. At first sight, the two types of fields are not that very different, since they both exhibit a tensorial character. Yet, physically speaking, we want them to do two different jobs and therefore it would help if, mathematically speaking, they were also of a different nature. These two distinct functions of the two types of fields are unfolded in an exemplary way in the context of the fibre bundles. In this context, the material tensor fields of all types appear as what we will call 'cross-sections', while for the carriers of the interaction -or force fields- we can employ the so called connections, which are another type of objects dwelling in the fibre bundle 'zoo'.

3.4.2 Yang-Mills Theories: the Weak and the Strong

In the previous sections we discussed the case of electromagnetism and we saw how interactions arise when we use the notion of gauge symmetry. Electromagnetic interactions arise when we require the Lagrangian of the system to be invariant under local gauge trans-

[123] Even if it is only in the sense that using this theory we got good explanations and very successful experimental predictions.

formations. This type of transformations belong to a larger class of transformations that we call Abelian because the group of transformations involved is the Abelian group $U(1)$. All the other fundamental interactions we know that occur in nature are described also by gauge symmetries, only these are more complicated since the groups involved are non-Abelian. These theories are also known as Yang-Mills theories because the first ones to employ them in the form they are known nowadays were Yang and Mills in their 1954 paper[124]. The only fundamental interaction that seems to be somewhat different is the gravitational, but we are not concerned with this issue in this thesis. In what follows we will concentrate on the Yang-Mills theories that are used in the description of the weak and the strong interactions, which employ the $SU(2)$ and the $SU(3)$ groups respectively, and once again we will only discuss the main ideas behind the formalism, rather than presenting it fully[125].

As we have seen, the starting point for the description of electromagnetic interactions was the observation of the invariance of the Lagrangian under global phase transformations $\varphi \rightarrow e^{-i\Lambda}\varphi$ of the wavefunction. By rendering the transformations local, $\varphi \rightarrow e^{-i\Lambda(\mathbf{x},t)}\varphi$, from the transformation requirements of the gauge fields, $A_\mu \rightarrow A_\mu + \partial_\mu \Lambda(\mathbf{x},t)$,we got coupling terms that allowed for the description of interactions. In the case of weak interactions we follow a similar process, but here the matter fields are multiplets, rather than scalars, and hence the transformation operators take the form of matrices, while the transformation parameters or gauge fields are vectors is some internal space. Hence, the trans-

[124] We have already seen in chapter 1 of this thesis that Klein anticipated Yang-Mills theories by fifteen years Utiyama discovered them independently and almost simultaneously with them and Shaw developed something similar right after them (1955). Nevertheless, Klein's work does not go as far as the work of Yang and Mills and Utiyama publicized his own a year after Yang and Mills. For more on the issue see O'Raifeartaigh, *The Dawning of Gauge Theory*.

[125] For detailed analysis see Aitchison & Hey, *Gauge Theories in Particle Physics*, or Ryder, *Quantum Field Theory*, or Balin & Love, *Introduction to Gauge Field Theory*.

formation for the matter field takes the form

$$\varphi \rightarrow \varphi' = \mathbf{U}\varphi$$

where $\mathbf{U}$ is a unitary matrix $n \times n$. For a scalar field and $n = 1$ we get the Abelian case we have already examined. Putting the transformation in exponential form we get the expression

$$\varphi = \exp(\frac{i}{2}\boldsymbol{\alpha} \cdot \boldsymbol{\tau})\varphi'$$

where $\boldsymbol{\alpha}$ represent the transformation variables and $\boldsymbol{\tau}$ are the generators of the (infinitesimal) $SU(2)$ transformations. When the transformation variables acquire spatiotemporal dependence, the gauge transformations become local and take the form

$$\varphi = \exp(\frac{i}{2}\boldsymbol{\alpha}(x) \cdot \boldsymbol{\tau})\varphi'.$$

To restore invariance of the Lagrangian under local gauge transformation we have to modify the transformation rules for the derivative as follows:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + \frac{i}{2}\boldsymbol{\tau} \cdot \mathbf{W}_\mu(x)$$

where ∂_μ is understood to be multiplied by an $n \times n$ matrix and the $\mathbf{W}_\mu(x)$ are three independent gauge fields $\mathbf{W}_\mu = (W_\mu^1, W_\mu^2, W_\mu^3)$ when we deal with the $SU(2)$ group which describes weak interactions. These are the generalization of the A_μ electromagnetic field and they are called Yang-Mills fields. The non-Abelian character of the Yang-Mills fields is displayed by the commutation relations

$$[\boldsymbol{\tau}_i, \boldsymbol{\tau}_j] = i\varepsilon_{ijk}\boldsymbol{\tau}_\kappa$$

that do not exist in the Abelian $U(1)$ case of the electromagnetic field.

Infinitesimal $SU(2)$ transformations take φ to

$$\varphi' = (1 + \frac{i}{2}\boldsymbol{\tau} \cdot \boldsymbol{\eta}(x))\varphi$$

and $\mathbf{W}_\mu$ to

$$\mathbf{W}'_\mu = \mathbf{W}_\mu - \partial_\mu \boldsymbol{\eta}(x) - (\boldsymbol{\eta}(x) \cdot \mathbf{W}_\mu).$$

As we can see, the first term in the transformation law for the gauge field is the generalization of the electromagnetic case. The second term reveals the fact that the three components of $\mathbf{W}_\mu$ form the components of a triplet representation of $SU(2)$. The covariant derivative D_μ of the matter field transforms in the same way as the field itself, namely

$$D'_\mu \varphi' = (1 + \frac{i}{2}\boldsymbol{\tau} \cdot \boldsymbol{\eta}(x)) D_\mu \varphi$$

and this restores invariance in the Lagrangian which before this modification had the general form $\mathfrak{L} = \mathfrak{L}(\varphi, \mathbf{W}, \partial_\mu \varphi, \partial_\mu \mathbf{W})$.

The strong interaction terms arise when we require invariance of the Lagrangian under $SU(3)$ symmetry transformations, in which case the above generalize as follows[126]. The $SU(3)$ group has eight generators, which means that the matter fields transform according to the law

$$\varphi' = \exp(i\mathbf{G} \cdot \boldsymbol{\alpha})\varphi.$$

The scalar product in the exponential involves 8-component vectors, $\mathbf{G}$ for the generators and $\boldsymbol{\alpha}$ for the transformation variables. Once again, the generators do not commute, instead they satisfy an algebra of the form

$$[G_i, G_j] = ic_{ijk}G_k.$$

[126] Here we follow Aitchison & Hey.

To make the above transformation local, we introduce eight gauge fields $W_\mu^1, ..., W_\mu^8$ and define the covariant derivative

$$D_\mu = \partial_\mu + i\mathbf{G} \cdot \mathbf{W}(x)$$

where the $\mathbf{G}$ are some set of matrices of appropriate dimension to act on φ. The infinitesimal transformation law for the Yang-Mills gauge fields is then

$$W_\mu^i \rightarrow W_\mu^{i\prime} = W_\mu^i - \partial_\mu \eta_i(x) - c^{ijk}\eta^j W_k^\mu(x)$$

where we can see, once again, that the first term of the transformed Yang-Mills field is the generalization of the electromagnetic case, while the second term tells us that the eight gauge fields $\mathbf{W}_\mu$ transform in such a way that the transformation coefficients are the structure constants of the group; because of the way they transform, we say that they belong to the regular representation of the group.

The difference of transformation laws between the non-Abelian and the Abelian gauge fields results in self-coupling terms in the Lagrangians of the former. In other words, in the non-Abelian case the Yang-Mills fields interact with themselves. Hence, a non-Abelian gauge system without matter fields has non-trivial interactions and therefore it is not free. This means that, basically, the gauge fields correspond to physical interactive entities in a straightforward manner. Unlike the Abelian case where the status of the gauge field is dubious and object of a major debate, as we shall see in the next chapter, in the case of the weak and the strong interactions the currents that are associated with the gauge fields are measurable and in this sense existing fundamental entities that interact directly with either matter fields or with each other. Hence, in the case of the weak and the strong inter-

actions, the surplus structure of the electromagnetic gauge theory becomes mathematical structure with elements corresponding to elements of the physical system.

The analysis above is just a very brief summary of what the generalization of Abelian gauge theories looks like. To do justice to the theory one would have to study it in detail and discuss the notions of symmetry breaking and asymptotic freedom, both necessary in order for the correspondence between the mathematical and the physical to be understood fully. But for the purposes of this thesis, this rather sketchy presentation suffices.

One thing that we would like to mention here is that the above formalism provides a counterexample to Field's programme. In chapter two, we presented Field's programme and Shapiro's criticism of it. Shapiro proved that it is possible to find an expression that is derivable from the supposedly conservative extension of the theory and yet it belongs to the actual physical part of it. Our view is that in the weak and the strong interactions, the gauge fields themselves exemplify such a case. Assuming that a nominalist is able to overcome Malament's objections and define congruence and betweenness that would allow for a full field theory to be expressed in a nominalistic way involving a mathematical and a physical part, one should be able to dispense with the conservative extension of the theory and derive all the physically significant results from its theoretical part only. The gauge fields live as cross-sections in the mathematical structure called the principal bundle and although they, themselves, are not just mathematical artifacts that one could dispense with[127], they are derivable from what could be considered to be the extension of the mathematical formalism only. The reason is that they emerge only if we consider that

[127] At least, the gauge fields would qualify as theoretical entities, in Field's terms, and theoretical entities are not dispensable.

there is a symmetry group in operation, namely the $SU(2)$ group for weak interactions. Then, given the gauge freedom of the theory, the connections -or gauge fields- are derived by the mere requirement that the theory is covariant under local $SU(2)$ transformations. There is no way of anticipating the existence of gauge fields and of their corresponding currents from the 'physical' part of the theory only. Yet, when we study weak interactions experimentally, the gauge fields appear to be as physical as any other interacting field; not only they 'click' but they also interact weakly or even electromagnetically[128]. In M. Redhead's terminology, the connections could be said to 'move across' the surplus structure boundary, thus descending from the mathematical to the physical realm.

3.5 Constrained Hamiltonian Systems or Fibre Bundles?

It is a common place view in physics that physical objects interact and it is through their interactions that we observe them. Therefore we need a description that accounts for these interactions and explains our observations. One very fruitful[129] way of describing interactions is by using variational calculus and local symmetries. So far in this chapter, we have become acquainted with the notion of symmetry as this occurs in the context of constrained Hamiltonian systems and we have already shown how symmetries allow for the description of interactions. But aside from this, or rather subsequent to it, there is another more elaborate formalism, the fibre bundle formalism which, we will argue, is more appropriate for describing interactive and interacting fields.

---

[128] Two out of the three carriers of the weak force have electromagnetic charge as well.

[129] Fruitfulness from this perspective means that it has given good descriptions/explanations and accurate predictions.

The Hamiltonian systems with gauge symmetries -Abelian or Yang-Mills- are constrained Hamiltonian systems. In their present form they first appeared in the Yang-Mills 1954 paper, as we have seen, and they came to the forefront of research in physics from the late 1960's onwards. In the mean time, since the 1930's, mathematicians who were studying relations between topology and geometry, and then from the 1950s onwards topologically non-trivial manifolds, developed the so called fibre bundle formalism, a generic geometrical approach that encompasses the mathematical structures that describe systems with constraints imposed by gauge symmetries. Fibre bundles were explicitly utilized in the formulation of gauge theories for the first time by Wu and Yang (1975), who compiled a 'dictionary' translating between the physicist's terminology and the new mathematical terminology. Here we have one more example of mathematical structures that develop regardless of the needs of the physicists' community, which find applications in physics later on. As we have already seen in the first chapter, in this particular historical incident a crucial idea that was (one of) the main motivations for the programme was common in physics and in mathematics. The reason why the development of the physical theory was slower than that of the mathematical, we argued, was the fact that there had not been much support from the experimentalists' front for a few decades. On the other hand, mathematicians who do not need phenomenological pick-ups to motivate their research proceeded immediately after the first ideas were presented and hence got there first.

Our aim in this section is to comprehend how systems with gauge symmetries are described in this formalism and what are the advantages of it when compared with the

constrained Hamiltonian formalism we introduced previously. What is more, we will try to do this using no mathematics at all.

3.5.1 Explaining Fibre Bundles

Is it possible to understand the fibre bundle formalism without using loads of mathematics? A simple answer to this question is 'no'! It is known since the times of Euclid's that there is no royal way to geometry, and things have changed little since then. For someone to understand and appreciate the fibre bundle formalism fully, one has to study it thoroughly, because it is only through study that one gets clear insights into certain geometrical concepts. In my view, this understanding comes in a non-verbal way and it is therefore rather difficult to put in words. But what I am hoping to do in the first part of this section is to give a description of the concepts involved and, where possible, to illustrate them by giving examples that are fairly easy to visualize, thus developing a pictorial understanding of some parts of the formalism. Then, one may be able to extend those intuitive images and complete the picture as much as possible, always bearing in mind that this is not the whole story, nor the correct/true one. Nevertheless, let us try to do this.

What a Fibre Bundle Is

Fibre bundles are a generalization of the Cartesian product in the following sense. A fibre bundle is a triplet $(\mathcal{M}, \pi, E)$ where $\mathcal{M}$ is what we call the base manifold, E is the total space and π is a projection map $\pi : E \longmapsto \mathcal{M}$. The inverse image π^{-1} of the map π takes you from a point $x \in \mathcal{M}$ to E and it is called the fibre $F := \pi^{-1}(\{x\})$ over x. The total space E is $\mathcal{M}$ itself along with the bundle of all the fibres over all

$x \in \mathcal{M}$, or $E := \cup_{x\in\mathcal{M}} F_x$. In certain cases, the total space E is the product space $\mathcal{M} \times F$ which is a generalization of the Cartesian product indeed. As we know, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are differentiable manifolds, then $\mathcal{M}_1 \times \mathcal{M}_2$ can be given a manifold structure where $\dim(\mathcal{M}_1 \times \mathcal{M}_2) = \dim(\mathcal{M}_1) + \dim(\mathcal{M}_2)$. But in fibre bundles the total space is not, in general, a product space and this will be made clear by two illustrative examples.

The first example is that of the product bundle[130]. The product bundle is one of the simplest examples of a fibre bundle and its three elements are: $\mathcal{M}$, $\pi = pr_1$ is the projection map taking you from any point of F_x, the fibre over x, to the point x on the manifold, and $E = \mathcal{M} \times F$.

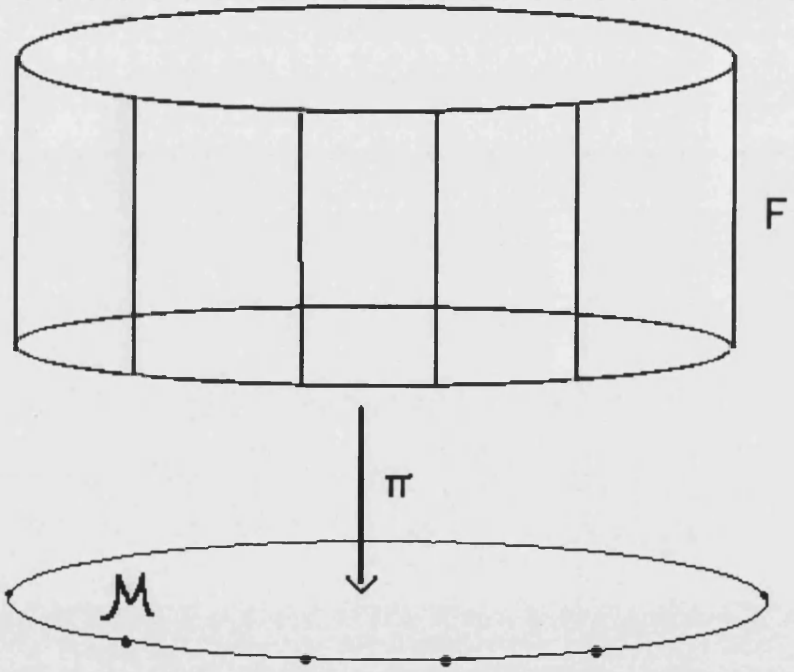


Figure 5
Product Bundle

Another example of a fibre bundle is the Möbius strip. Here, the base space $\mathcal{M}$ is the circle $S_1$and the fibre could be taken to be the interval $[-1, 1]$. But the total space E is not the product space $\mathcal{M} \times [-1, 1]$, nor is it homeomorphic to it because the total space is

130 For more details see C. Isham, *Modern differential Geometry for Physicists*, pp.204-6.

twisted. It can be represented, instead, by a rectangle whose short edges identify as shown in the picture.

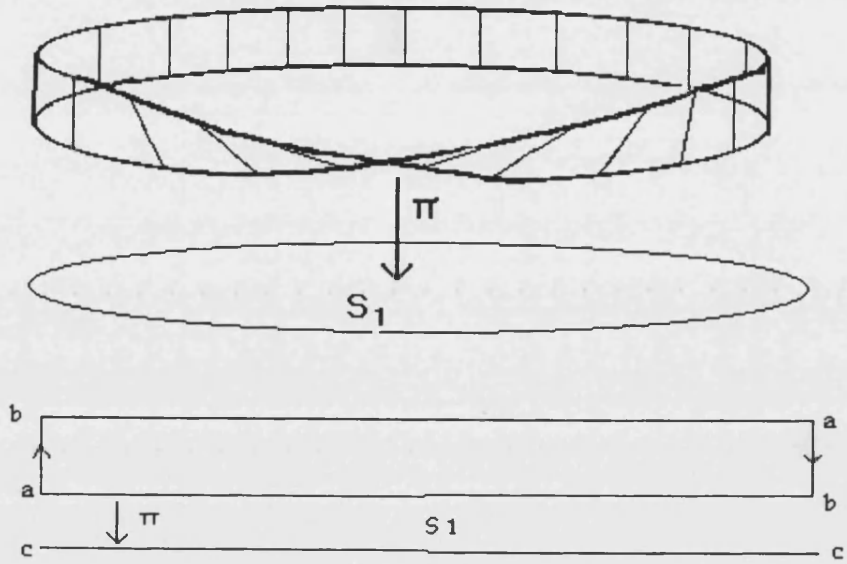


Figure 6
The Möbius Strip

Cross-Sections

The notion of cross-section is very crucial in both the fibre-bundle formalism and its application in physics, since all the matter fields are defined as cross-sections of the tangent bundle; the tangent bundle, a special case of a fibre bundle, we will be discussing in the next section. The cross-section is a map $s : \mathcal{M} \longmapsto E$ such that the image of each point $x \in \mathcal{M}$ lies in $\pi^{-1}(\{x\})$. π and s are inverse to each other:

$$\pi \circ s = id_{\mathcal{M}}$$

So, here we are talking about some mathematical object (a field) which takes some *specific* values across the fibres as its location on the base manifold changes too. So far as

the product bundle is concerned, the cross-section is defined uniquely and continuously everywhere.

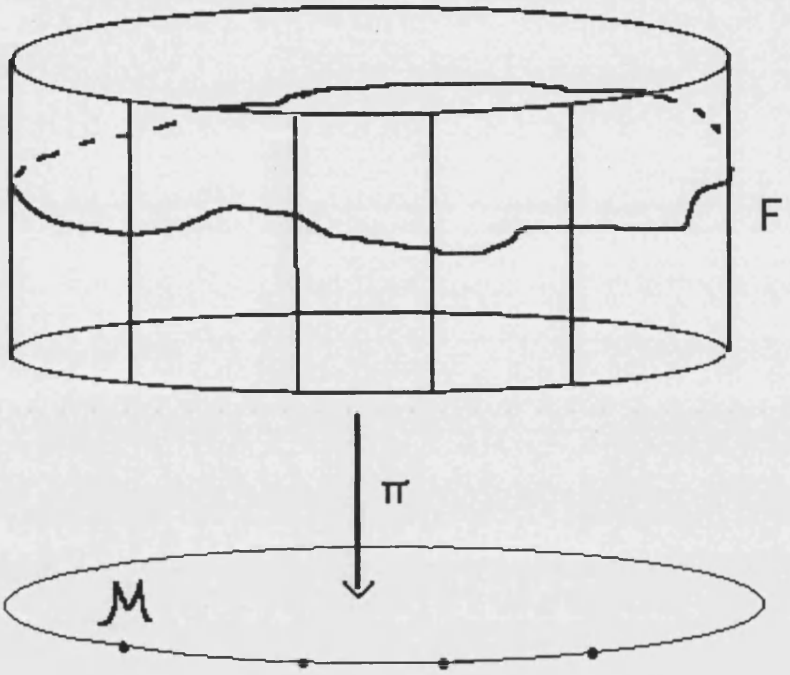


Figure 7
Cross Section of a Product Bundle

But in the case of the Möbius-strip-bundle, which is a non-orientable surface, this is not the case as we can see.

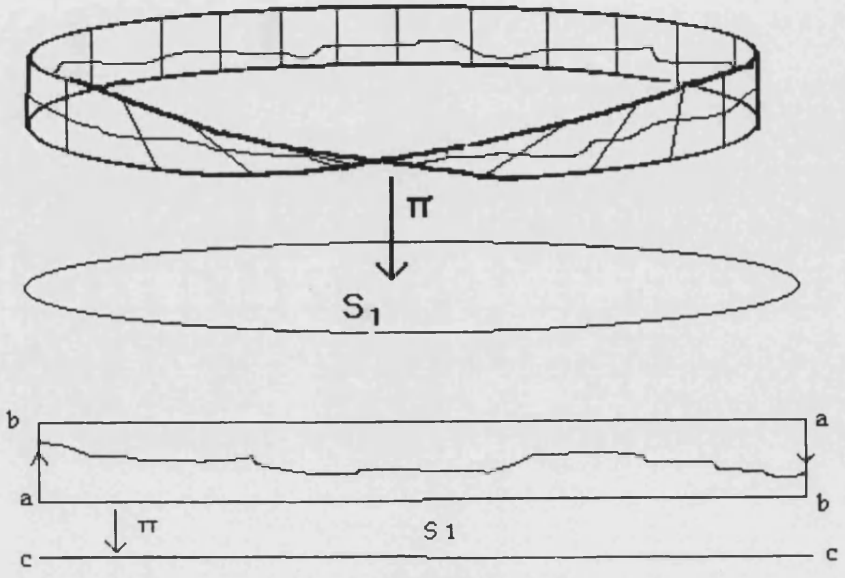


Figure 8
A Cross Section of a Möbius Bundle

Here it becomes obvious from the picture that the cross-section is not continuous, as we can see from the figure above. In other words, the cross-section is equivalent to a function from S^1 to $[-1, 1]$ which is antiperiodic around the (circle) base manifold. The Möbius strip is just an example of a non-orientable fibre-bundle, but from that we can see how the cross-section and its continuity depend upon the topology of the total space. At this point we have to make a leap. In general, in the cases of the so called principal fibre bundles where the bundles have the special structure of a vector space, the following theorem holds.

Theorem 4 ***A principal fibre bundle has a continuous cross-section if and only if it is trivial***[131].

One of the two things this theorem tells us is that when the topology of the base manifold is non-trivial, we will not find a continuous cross-section. So, if we take the base manifold to represent spacetime, then if the topology there is not trivial, we are not able to define vector fields continuously all over it, and this, as we shall see, is related to the well known problem in gauge theories, the so-called Gribov obstruction, which does not allow us to determine the gauge everywhere at once. But on this point, more discussion follows later in this chapter.

Principal Bundles, Vector Bundles and Connections

At this point we need to make another leap and try to visualize two more complicated examples of fibre bundles, having as a starting point the simple cases of the product and the Möbius bundle. The first case is that of the tangent bundle, which is the bundle of the

[131] For a proof of this theorem, see C. Isham, Modern Differential Geometry for Physicists, 2nd ed., p.230.

tangent spaces at all points of a base manifold, while the second is the bundle of frames, which, as its name indicates, is the bundle of all frames at all points on the base manifold. In order to get a visual idea of what the various objects involved represent, we will use the following illustration[132].

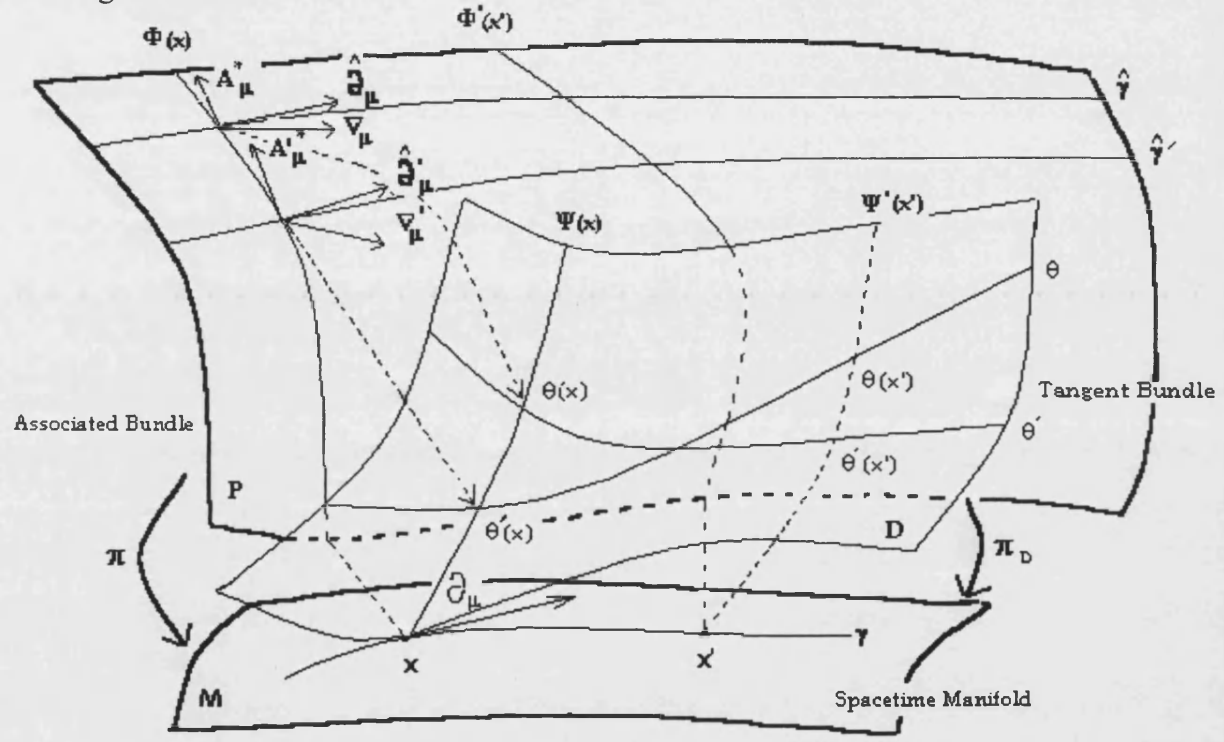


Figure 9
Associated and Tangent Bundles

The Tangent Bundle: a Special Example of a Vector Bundle

The base space $\mathcal{M}$ of the tangent bundle may be considered as the $4-$dim spacetime manifold. The fibre F_x over each point x of the manifold is the tangent space $T_x\mathcal{M}$ to $\mathcal{M}$ at the point x which is generated by all the tangent vectors at this point; or in other words, by the vectors of all the curves which pass through the point x and are tangent to x. The total space E , or the tangent bundle $T\mathcal{M}$, is defined as $T\mathcal{M} = \cup_{x\in\mathcal{M}} T_x\mathcal{M}$, the union of

[132] This illustration is based on the figure B3 of p.220 of Sunny Auyang's "*How is Quantum Filed Theory Possible?*"

all tangent spaces at all points of the manifold $\mathcal{M}$. Each fibre F_x, or in this case $T_x\mathcal{M}$, is nothing other than *the set of all vectors that are tangent to the manifold at that point.* For each tangent space , the following theorem holds.

Theorem 5 *The tangent space $T_x\mathcal{M}$ carries a structure of a real vector space.*

It can also be shown[133] that the tangent bundle $T\mathcal{M}$ has a natural structure of a $2m$-dimensional differentiable manifold, where m is the dimension of the manifold $\mathcal{M}$ itself.

The cross-sections θ of this vector bundle are used for the description of matter fields with phase θ. Along each cross-section, the wavefunction of the matter field may take different values, but its phase remains the same, i.e. $\vartheta(x)$. The information encoded here is that as we move along a curve γ on the base manifold, the phase of the field may or may not change and this depends on the interactions which may be accounted for by the connections, as we shall see shortly.

The Bundle of Frames: a Special Example of a Principal Fibre Bundle

A more complicated case of a fibre bundle is the bundle of frames, which is a special case of what mathematicians call a principal bundle. A principal fibre bundle is one whose fibres are Lie groups in a specific way. The principal fibre bundles "have the important property that all non-principal bundles are *associated* with an underlying principal bundle. Furthermore, the twists in a bundle associated with a particular principal bundle are uniquely determined by the twists in the latter, and hence the topological implications

[133] See C. Isham, p.89.

of fibre bundle theory are essentially coded into the theory of principal fibre bundles"[134]. A typical example of a principal fibre bundle is the bundle of frames.

In the case of the bundle of frames, the base space $\mathcal{M}$ is, once again, an m-dimensional differentiable manifold which we may consider to be the $4-dim$ spacetime manifold. A linear frame, or base, at the point $x \in \mathcal{M}$ is an ordered set $(b_1, b_2, ..., b_m)$ of basis vectors for the tangent space $T_x\mathcal{M}$. In this case, the projection map $\pi : \mathbf{B}(\mathcal{M}) \rightarrow \mathcal{M}$ is defined to be the function that takes a frame into the point x in $\mathcal{M}$ to which it is attached. The fibre over $x \in \mathcal{M}$ is, of course, the inverse image under the map π and it comprises the set of all the local frames that are associated with the point $x \in \mathcal{M}$. The total space of the bundle of frames, which we denote by $\mathbf{B}(\mathcal{M})$, is the *set* of *all frames* at *all points* of $\mathcal{M}$. $\mathbf{B}(\mathcal{M})$ is a right G-space, where the group acting on it is the $GL(m, \mathbb{R})$, as well as a differentiable manifold of dimension $m + m^2$.

In our graphic representation of the principal fibre bundle we can see the following. 'Over' each point x of the base space $\mathcal{M}$ there is the fibre of x, represented as a line with $\phi(x)$ at the top. The cross-sections of this fibre bundle are depicted by the $\hat{\gamma}$-lines and they introduce a specific coordinate system along the curve γ so that as we are moving along the curve we have a fixed coordinate system or frame -this could be understood as an active transformation where the actual system is 'moving' but the frame remains the same. As we move along the fibre, the value of the field ϕ does not change but the frames do -this is what we could understand as a passive transformation where the physical system remains fixed but its description changes.

---

[134] C. Isham, *Modern Differential Geometry for Physicists*, 2nd edition, p.220.

If, instead of the bundle of frames, we had chosen a fibre bundle with symmetry group the $SU(2)$, we would have the $SU(2)$-bundle of the Yang-Mills theory. In this case, selecting a specific cross section is also known as gauge choice or gauge fixing.

Connection on the Bundles or Moving Around

Next, we need the notions of the connection and of the pull-back. The connection tells us all about how we move around in the bundle, while the pull–back is the operation we need in order to be able to 'move' from the total to the base space and the other way round.

The connection is a field defined on the bundle space and, as its name indicates, basically we need it so that we can connect or compare points in 'neighboring' fibres in a way that is not dependent on any particular local bundle trivialization (i.e. choice of frame). This suggests that we should look for vector fields on the bundle space P that 'point' from one fibre to another[135]. What is needed, therefore, is some way of constructing vectors that point away from the fibre, i.e., elements of T_pP that complement the vertical vectors in V_pP.

In general, in the bundle of frames, the symmetry transformations are diffeomorphisms on the bundle space. These transformations we could view in two ways, active or passive. Active transformations take the point x of the manifold $\mathcal{M}$ to the x', while the passive transformations change things on the bundle space but leave x unaltered, so that the only thing that changes is the coordinate patches. One may then ask: and what can we actually do with the connections? Well, in the active case, and while still on the bun-

[135] See C. Isham, p.253 for a more detailed discussion.

dle of frames, the connections describe how the field of frames changes as we move along a spacetime path and therefore 'hop' from one fibre onto another. As a physical system moves along a spacetime curve γ, the tangent spaces change and so do the frame-fibres. In general, these tangent spaces are not in any natural relation to each other. The connection, represented by ∇_μ, allows us to compare these spaces, by expressing how $\hat{\partial}_\mu$ changes as we 'cross' different bundles. If the local representative of the connection was given the name A^*_μ[136], this could be represented in a diagram as follows:

$$\begin{array}{ccc} A^*_\mu & - - - > & A'^*_\mu \\ \uparrow & & \uparrow \\ x & - - - > & x' \end{array}$$

All change is determined by the connection but, as we should expect, this is done in a non-deterministic way; if there is no necessity to impose a choice of a specific cross-section, the evolved system may start from any point of the initial fibre and be found on any point of the final. However, as we can see from our illustration, when moving along γ and at the same time staying on the same cross-section, the initial coordination remains the same; which means that we know exactly where we will find our system when we are looking for it in the total space.

The passive view of the transformation is somewhat more difficult to describe correctly here, because the actual illustration is inaccurate and incomplete[137]; but the intuitive

---

[136] As a matter of fact, the connection is usually associated with a certain $L(G)$-valued one-form ω on the bundle space P, while by Γ we denote the associated $L(GL(m,\mathbb{R}))$-valued one-form on $U \subset \mathcal{M}$ and the symbol A^α_μ is used specifically for the Yang-Mills field, which can be regarded as a Lie-algebra valued one-form on $\mathcal{M}$, at least locally. In this paper, we chose to use the symbol A^*_μ for simplicity and to give some sort of unity. I would like to make it clear, though, that this 'unified' use of one symbol is not accurate and I would like to warn the reader that this may be confusing if they study, for example, C. Isham's book.

[137] For more extended discussion see C.Isham (1999).

idea is the following. For the description of the same spacetime point, we may use more than one different coordinations, which are related to each other by the action of the group $GL(m, \mathbb{R})$. Thus the 'location' on the bundle space, or the local trivialization, changes, while the physical system remains where it was in the spacetime manifold. In this case, the connections corresponding to the two different local trivializations are the transform of each other under the action of the group. In the form of a diagram, the situation could be illustrated as follows:

$$\begin{array}{ccc} A^*_\mu & - - - > & A'^*_\mu \\ & \nwarrow \swarrow & \\ & x & \end{array}$$

In general relativity, the role of the connection is played by the well-known Christoffel symbols. In Yang-Mills theories, on the other hand, where the principal bundle is one with a Lie group acting on it, the role of the connection is played by the Yang-Mills field itself.

Gauge Transformations

If we want to be more accurate, we have to say that the connection is an $L(G)$-valued one-form on a principal bundle[138] and it is such that it can be decomposed locally as the sum of a Yang-Mills field on $\mathcal{M}$ plus a fixed $L(G)$-valued one-form on G. Since the latter $L(G)$-valued one-form is fixed when we know the Yang-Mills field, basically, we know the connection, at least locally. So, in this informal sense, we could 'identify' the connection with a Yang-Mills field -as we have done above. What we need to look at here is how

[138] For a detailed discussion see, for example, C. Isham (1999), pp.254-262.

gauge transformations come up in fibre bundles and how the constraints and hence how the conserved currents are represented.

In general, a gauge transformation is considered to be *any* automorphism of the bundle. In the case of passive transformations the actual transformation map $\phi : P \rightarrow P$ takes you from a coordinate chart to another -the two have overlapping domains U and U'. Then, it can be shown that the transformed connection is also a connection and that the transformation of the local representatives of the connection, i.e. of the Yang-Mills field, is our familiar gauge transformation. Along the same lines, when we consider active automorphisms of the bundle, the transformation on the bundle induces a transformation of the connection that locally is exactly like the familiar gauge transformation of the gauge field; the only difference here is that the diffeomorphism is defined on the manifold $\mathcal{M}$ as $h : \mathcal{M} \rightarrow \mathcal{M}$.

Mathematically speaking, the two different ways of viewing transformations, aka the active and the passive, are equivalent. Yet, when we use this formalism to represent physical structures a problem arises. The active transformation is considered to correspond to actual transportation of the physical system from one spacetime region to another. The passive transformation, on the other hand, changes only the description of the system, the coordination one could say. In what sense, then, are the two equivalent when we talk physics? If we claimed that a transformation/change in the description of a structure corresponds to an altogether new 'reality' in a sense, similar to that of a physical structure that has been transported to a new spacetime region, would we do justice to the mathematical equivalence? Or is this a far fetched assumption? Because in the active case, there is

some actual change of the physical structure we study, but in the passive case there does not seem to be any. Except, if some of the mathematical objects that live in the bundle space and undergo a change, the connections for example, did correspond to physical structures. If that was the case, we could comprehend how the two types of transformation are equivalent in a physical sense. But the question of whether the connections have physical status is one to which we cannot give a straightforward answer, at least not right now, because although we make use of the connections to represent the interactive fields, we cannot say before we give it some further thought that these are indeed 'tangible' physical objects. Note in passing that the same sort of question is addressed by Redhead (2001) who claims that when the automorphisms of the physical and the mathematical structure are in one-one correspondence and since the symmetries of the physical structure express important structural properties of it, so would do the symmetries of the mathematical structure. Things are somewhat different, though, when symmetries are present in the surplus structure, in which case the mathematical symmetry gives interaction terms in the physical structure. It remains to show how this relation between the two manifests itself , but we cannot do this before we investigate the role of the connection in the description and explanation of certain physical processes; so, we will try to answer this question later on, mainly in the following chapter.

Finally, let us turn now to the idea of the constraint, as this may be understood in the fibre bundle context. In Hamiltonian systems where symmetry transformations leave their action unaltered, we get, according to Noether's theorems, conservation laws and constraints. The conservation laws, as we have seen, involve derivatives of the fields involved

and hence they impose the symmetry conditions that define the bundle space. Hence, we could understand the constraints as restrictions that are imposed on the evolution of our original system and on its 'behavior' in the bundle space, or in other words as the gauge orbits.

Associations

The tangent bundle and the principal fibre bundle that can be seen on the illustration, are associated bundles. In general, the basic intuition that underlies their association is that "given a particular principal bundle $(P, \pi, \mathcal{M})$ with structure group G, we can form a fibre bundle with fibre F for each space F on which G acts as a group of transformations"[139]. In our specific example, the group of the bundle of frames $GL(m, \mathbb{R})$ acts on the tangent space on each point of $\mathcal{M}$ and the result of the action is the change of the mathematical expression of the local coordinate chart in a passive way, if x does not change and therefore we are still on the same fibre, or in an active way, when x changes as well. So, for the same x, a symmetry transformation could take the connection field $A^*_\mu(x)$ to $A'^*_\mu(x)$, while an active one could take it to $A^*_\mu(x')$ or $A'^*_\mu(x')$ depending on whether we stayed on the same cross-section or not. These changes on the principal bundle are linked with changes on the associated tangent bundle in the following way. When we are considering passive transformations, the action of $GL(m, \mathbb{R})$ on the vector space of the tangent bundle can be understood as changing the direction of the tangent vector on x, while still remaining on the same tangent 'plane' or fibre; so it takes you from $\psi(x)$ to $\psi'(x)$. When the transformations are active, there is a total change of the ψ-field -i.e. change which affects both the spacetime

[139] Isham (1999) p.232.

point and the fibre. So, if the transformation leaves the field on the same cross-section, the transformed field will be $\psi(x')$ while if not on the same cross-section, the transformed field will be $\psi'(x')$.

When the group acting on the principal bundle is a gauge group, the action of the group on the associated bundle will be expressed as a change of the phase of the matter field -with or without simultaneous change of its spacetime location, depending on whether the transformation is considered as active of passive respectively. In the active case, starting with phase $\vartheta(x)$, we end up to one with phase $\vartheta(x')$. On the other hand, an active action of the group projects the original point of the total space to some other point which lies on a different fibre altogether. In this case, if we are still on the same cross-section, the transformed phase will be $\vartheta(x')$, while if we are not, the new phase will be $\vartheta'(x')$.

This association between principal and vector bundles is what allows coupling terms to appear; it is precisely these terms that can be interpreted as interaction terms when we are using the formalism to describe interactive fields in field theories.

In concluding this section we need to address an important question. If we should take realistically one of the two spaces, namely E and $\mathcal{M}$, what should we consider as physically real, the spacetime manifold or the total bundle? This is an issue similar to one that has already been addressed in the context of general relativity and is known as substantivalism. I am leaving the question unanswered for the moment and we will get back to it later on .

3.5.2 Science With Numbers, but not Necessarily With Coordinates

One further advantage of using Fibre Bundles is that the formalism is such that we do not need any reference to any kind of coordinates and reference frames. We are enabled, therefore, to express the laws in a coordinate-free way and thus to have them in their most general form.

For example, instead of the familiar form of Maxwell's equations in classical physics which is coordinate dependent, using exterior calculus we may formulate them in an intrinsic, coordinate-free way. So, Maxwell's first and fourth equations $curlE = -\frac{1}{c}\frac{\partial B}{\partial t}$ and $divB = 0$ become $d\eta = 0$[140].

In the previous chapter we discussed Field's objection to using numbers and his suggestion to consider spacetime points as the fundamental entities of physics. His idea was that we could consider spacetime points, instead of numbers, as fundamental entities and attribute properties to them and therefore account for everything happening using mathematics as a conservative extension of the physical theories. We also mentioned there that Field favored the use of tensor calculus because by employing tensors one does not have to appeal to numbers; the drawback of using tensors, though, is that one does not avoid the use of scalar magnitudes that may be chosen arbitrarily and hence his own nominalistic approach does better than tensors in avoiding arbitrary choices.

After the discussion in this chapter it has become clear, we suppose, that, first of all, in order to describe interactions we need two different types of entities acting together at the same regions of spacetime points. Hence, according to Field's programme we would have

---

[140] See also Darling, *Differential Forms and Connections*.

to ascribe to the same physical entity two different bunches of properties, which are not the same as ascribing, say, extension and temperature. For, while in the case of extension and temperature we would just attribute two different properties to the same entity, in this case we would have to impose on the same object the characteristics of an interactive entity and of the interacted one at the same time. Hence, in our view, by doing something like that we basically remove the possibility to account for distinct physical objects whose existence has been verified experimentally and to give causal explanations. In other words, we should not be able to do physics any more.

So, in a nutshell, what we are trying to say is this. In quantum field theories, interactions are essential, since it is through them that we observe the physical structures. From the physics literature we see that gauge theories can describe field theories and interactions in them successfully. Interactions arise naturally as the solution of a variational problem. Fields carry tensor as well as spatiotemporal specifications to account both for 'where &when' (on the manifold) as well as for 'interactions'. Tensor fields spontaneously arise as cross-sections in fibre bundle theories, while the force fields are identified with the connections and this happens in a deductive top-bottom way. Using differential geometry -and more specifically, the fibre bundle approach- we may express interactions in a coordinate-free way, which is important because then they do not depend on any specific system of reference. For all the above reasons, it is obvious that differential geometry and the fibre bundles formalism are a 'natural' environment for gauge theories to flourish. They provide the most appropriate and agreeable formalism at present and we might even claim

that it is also a necessary one[141]. Moreover, the interaction fields behave differently from the matter fields -the former display a bosonic behavior (associated with integer spin) while the latter a fermionic one (which means half integer spin). It is clear that we do not really need any gauge principle in order to justify this approach. We do not need anywhere the claim that 'all fundamental interactions in nature obey a/the gauge principle' or that 'the gauge principle dictates the interactions'. To our view, what really happens is that the notion of gauge symmetries, rather than dictating to us how, it *enables* us to describe some specific types of interactions in a consistent, deductive way -a top-bottom approach, the holy grail of theoretical physics- and at the same time it allows us to investigate the possibility of describing all the fundamental forces in the same way. This is a whole research programme in its own right and it has proved a very successful one. Hence, one could claim that the gauge principle has been confirmed and established in an a posteriori way and we have to accept it as such, but not as a necessary principle imposed by nature.

[141] If there truly are in nature topologically non-trivial entities, then the fibre bundle formalism becomes indispensable. For more on this, see chapter 4.

Chapter 4
Scientific Explanation: Four Ways to the Aharonov-Bohm Effect

Up until now, we have discussed the relation between mathematics and physics and we have seen how some aspects of this relation are exemplified by quantum field theories when they are expressed in the form of constrained Hamiltonian systems; we have also illustrated how the same physical systems are described using a more elaborate tool, namely the fibre bundles formalism. Next we will examine more thoroughly the relation between this latter mathematical structure and the physical systems it represents in the context of the discussion of the second chapter and we will draw our attention to the advantages and the disadvantages of this formalism.

One of the major advantages of the fibre bundles formalism is that it provides a unified -in the sense of top-bottom- approach to the whole picture of interacting fields and hence it allows for what we will call holistic explanations of certain physical events; this is an aspect that the constrained Hamiltonian formalism fails to capture. Aiming to bring to light this advantage, in this chapter we will use as a case study the Aharonov-Bohm effect and after we look at three suggested explanations and the problems they encounter, we will examine a fourth approach. This kind of explanation does not clearly fit any of the models of scientific explanation set forth by philosophers and hence is a *sui generis* type worth examining in some detail. For this reason, we will begin this chapter discussing the notion of scientific explanation and the problems this concept encounters in philosophy of science

and then we will expand on this fourth explanation of the Aharonov-Bohm[142] effect, an explanation which uses the fibre bundles formalism.

4.1 Scientific Explanation

So far as scientific explanation is concerned, "the current situation is an embarrassment for the philosophy of science[143]" and it is so because although there have been several better or worse accounts about what scientific explanation is, there is still missing a single theory of explanation that could cover all possible examples. It may be the case, of course, that it is not viable to search for a single theory because there are, and always will be, scientific explanations of different kinds. Nevertheless, the purpose of this thesis is not to argue for or against the possible existence of a single theory of explanation. Instead, our intention is to examine the nature of a specific example of scientific explanation, test it out against the existing theories and evaluate its status with respect to those theories. Having this purpose in mind, we will run through the main proposals that are currently discussed and either endorsed or criticized by the philosophical community without trying to remedy their problems.

As one would expect, the classification of the approaches as to what scientific explanation is differ, according to various authors. But a reputable classification -and one that serves the purposes of this thesis as well- would be the very recent one by W. H. Newton-Smith (2000)[144]. There, he cites the following approaches.

---

142 Henceforth, we will refer to the Aharonov-Bohm effect as the A-B effect.

143 W. H. Newton-Smith (ed.), *A Companion to the Philosophy of Science*, Blackwell, 2000, p.132.

144 For detailed discussions on and different approaches to scientific explanation see, for example, Achinstein

First of all, is the so called *deductive-nomological*, or D-N, model of scientific explanation, introduced by Hempel. According to this model, a scientific explanation of a particular fact is nothing other than a deductive argument, where the premises comprise general laws as well as statements describing other particular facts and the conclusion follows from the premises. Such an argument is a scientific explanation just in case it is deductively valid. The main problem of this model of explanation is that it fails to accord with the fact that explanations are asymmetric, in the sense that when $\mathcal{A}$ explains $\mathcal{B}$, then $\mathcal{B}$ cannot explain $\mathcal{A}$.

An alternative to the D-N model of scientific explanation is the so called *causal-relevance* model, or C-R. This model emphasizes precisely the very fact of asymmetry and, according to it, explanations are no longer considered to be deductive arguments, but an account of the causal mechanisms that are responsible -partly or fully- for the phenomenon to be explained, the explanandum. The difficulties that this model faces are, first of all the fact that the notion of causation is at least as obscure and problematic as that of explanation itself, and second the fact that a great many of scientific explanations are not causal explanations, despite the fact that causal relations and factors may be involved.

Types of explanation that are not causal are explanations by *identification*, explanations using *models and analogies*, explanations *by unification* and explanations *focusing on pragmatic aspects*. In certain cases, the explanandum is explained by identifying some of its features with other observable facts and quantities that are better understood. For example, by identifying temperature with molecular motion, one can explain how the tem-

---

P., *The Nature of Explanation*, OUP, 1983, Cartwright N., *How the Laws of Physics Lie*, Clarendon Press, 1986, Ruben D.-H. (ed.), *Explanation*, OUP, 1994, Salmon W.C., *Causality and Explanation*, OUP, 1998.

perature of a gas increases when the average molecular speed increases as well. Our understanding of a complicated physical structure is improved when it is modeled by a simpler structure the workings of which we know. On the other hand, unification of, say, Newton's laws of motion and the universal law of gravity explains Kepler's laws of planetary motion by making them deductive consequences of a bigger structure[145]. This type of explanation is still open to further elaboration and refinement and its relations to the C-R model needs to be examined[146]. Finally, the view that focuses on pragmatic aspects takes into account the fact that the explanation which we would consider as satisfactory depends heavily on the context and good explanatory answers must be relevant. The problem with this last one is that the notion of relevance, as this was articulated by van Fraassen, is unconstrained and hence it virtually allows for anything to explain anything!

From what we can see so far, causation -never mind how problematic this notion may still be- plays quite an important role even in approaches to explanation that are not genuinely causal. So, in the cases of explanation by identification and of explanation using models, at some point or another one will appeal to causal factors that are involved. And even for the unification approach, Salmon (1998) has suggested that unification and C-R may be complementary rather than competing. In what follows, we will examine the relations between the two in the specific example of the A-B effect. Also, within the same

---

[145] The fact that a strict application of Newton's laws, applied to planetary models, must be amended by idealizations and approximations in order to yield Kepler's laws, strictly speaking, means that the deduction we are referring to above is not really a deduction. However, if we assumed the laws to be true -as we often do in physics- then Kepler's laws are deduced from Newton's.

[146] We will come back to explanations by analogy in the last chapter of this thesis.

context we will examine how well the specific explanation we have in mind fits with the D-N model.

4.1.1 Holistic *vs* Causal

In a somewhat different -as well as older- approach, Nagel (1961) distinguishes four types of scientific explanation: explanations that fall under the heading the *deductive model*, which is the same as the D-N model mentioned above, *probabilistic* explanations, *functional or teleological* explanations and *genetic* explanations. Probabilistic explanations are explanations that are definitely not of deductive form. In them, the explanans do not deductively imply the explanandum but they render it highly probable, or at any rate more probable than in the absence of explanans. Most statistical explanations in physics and in other sciences are of this type. For example, most of the explanations in nuclear physics and many in quantum mechanics could be considered to fall under this category. Genetic explanations, on the other hand, explain by describing the sequence of events that lead to the evolution of one system into another.

Finally, functional or teleological are characterized as the explanations that appeal to a final goal of the system we examine. Phrases that are common in such explanations are 'in order that' or 'for the sake of'. Nagel points out, though, that despite the common belief, teleological explanations are not necessarily anthropomorphic and that they do not demand that "the future is an agent in its own realization". And then he argues that although this kind of explanation is common in biological sciences, it is not exclusive to them for even in physics we do have explanations that share the main characteristics of teleological ex-

planations. The main examples he gives from physics are those of mechanical systems that employ the principle of least action and variational calculus. The systems that we have examined in the previous chapter are such systems, thus it is worth expanding on the notion of teleological explanation as this is explicated by Nagel and then, using the example of what we call the topological explanation to the A-B effect, relate this example to the teleological and compare it to the D-N and the C-R models of explanation. It is worth noting at this point that the explanation of the A-B effect we are offering here is not the only topological explanation that exists. Certain topological solutions of Yang-Mills theories are essentially topological but so are certain attempts to explain 'handedness' and projectile motion in classical mechanics. Postponing the discussion of all these topological explanations for later, let us now examine in some detail the notion of teleological explanation before we turn to the specific topological explanation of the A-B effect.

Teleological Explanations

Teleological explanations occur mainly in biology, as Nagel indicates, where processes are directed towards attaining certain end-products. Explanations in physics, on the other hand, are unlike the ones in biology since the notion of final cause is not considered at all in the study of physical phenomena. But then the question that arises is whether this disparity entails that there are no teleological explanations at all in physics and thus render biology an absolutely autonomous discipline. The answer he gives is 'no' and here is how he supports it.

First of all, he claims, teleological explanations are not equivalent to non-teleological ones. This can be seen easily when we consider first that although a teleological statement implies a non-teleological one[147], the inverse is not always true, therefore there must be some important difference between the two. So far as physical sciences are concerned, they do employ formulations that have at least the appearance of teleological statements, e.g. by using what he calls extremal principles -or the principles of least action, as are usually called. Principles of least action state that certain physical systems evolve so that their action, a magnitude from which all the possible configurations of a system are deduced, takes its smallest value. However, "such teleological interpretations of extremal principles are now almost universally recognized to be entirely gratuitous"[148] because even in physical systems obeying extremal principles there are no purposes or dynamic operations acting on their own right and directing the system towards a specified and specific goal. This lack of purposes is revealed by the fact that the dynamical structure of physical systems can be considered as the effect of constituent elements and contributory processes and not as the outcome of certain global properties of the system as a whole. The lack or the presence of global properties in a system taken as a whole will provide one of the ultimate distinctions between teleological and non-teleological explanations as they are usually enunciated. But before we elaborate on that, let us point out some more observations Nagel made about the differences between teleological and non-teleological systems.

---

147 For example, a teleological explanation of the fact that humans sweat when it is hot is that the human body maintains its temperature constant. A non-teleological explanation that follows from the teleological one is that when hot one puts on less cloths, seeks cooler spots, drink cold drinks etc. and all these help them maintain the temperature of the body constant.

148 Nagel, *The Structure of Science*, p.407.

In biology, usually we are concerned with a special class of organized bodies, like for example the pancreas, and we seek explanations about their functions which, in turn, lead us to investigate the conditions making for the persistence of this specific system. So, a statement of the type 'the secretion of insulin regulates the feeling of hunger so that the organism gets the food it needs for its maintenance' would constitute an explanation the explanatory power of which lies on the fact that there is a goal behind the response of the system: this kind of biological system responds to changes triggered by its environment by altering its functions so that its goal is sustained. The physical sciences, on the other hand, are not concerned with selected physical systems, nor with special classes of bodies. Instead they study the effects of certain conditions and processes on an unbounded variety of physical objects. Hence, when we study the radiation of the sun, for example, we may discuss its effects on a wide variety of physical systems and no such system is considered as more important. Moreover, there is no underlying goal in the systems and the processes concerned in this example: we do not 'explain' the average radiation per square meter on the surface of the earth on the basis of the maintenance of the average temperature of the earth. Nor do we claim that this quantity fluctuates according to the damage we -human beings- have done to the ozone layer so that the temperature of the planet and the amount of ultraviolet radiation arriving at its surface remain constant. This major difference between physical and biological systems, namely the fact that "living things exhibit in varying degrees adaptive and regulative structures and activities, while the physical systems

do not -so it is frequently claimed"[149] justifies the fact that teleological explanations seem "peculiarly appropriate" for biological but not for physical systems.

Yet physical systems that are self-regulating and self-maintaining have been constructed. Examples of such systems are automatic pilots, electronic calculators and thermostats, to mention just a few, and these systems resemble living organisms. So one may be justified to claim that there are non-vital systems that could be characterized as teleological and hence one needs criteria that would enable one to distinguish between them and non-teleological non-vital systems. Bearing in mind that physical scientists are justified to find objectionable the assumptions about underlying purposes in physical processes, we would be able to attribute a kind of 'goal-directedness' to physical as well as to biological systems only if it was possible to formulate the structure of 'goal-directed' physical systems in such a way that the analysis is neutral with respect to assumptions concerning the existence of purposes. This is possible, according to Nagel, when we characterize such systems as teleological on the basis of certain assumptions that render teleology into an analyzable category. The assumptions are the following: (*i*) the system S can be analyzed into a set of related parts or processes that are causally relevant, yet they can be assigned independently, to the occurrence of some property or mode of behavior G[150] of the system, (*ii*) a change (with time) in any of the variables that characterize the G state of S takes S out of this state; we call this change a *primary variation*, (*iii*) when a primary variation occurs in one or some of the parameters, the remaining parameters also vary so that they only

149 Ibid., p.408.

150 G contains in the form of variables -not necessarily numerical- all the independent parts of S that are causally relevant to the state of the system.

take values from certain classes of their range and we call this an *adaptive variation*, (*iv*) the values that the primary variation has assigned to the initially changed variables correspond to the values the adaptive variation has assigned to the adaptively changed variables so that S is eventually in a G state again. "When a system S satisfies all these assumptions for every pair of initial and subsequent instants in a time interval T, the parts of S causally relevant to G will be said to be 'directively organized' (during the interval T with respect to G)"[151]. This definition can now be used to characterize biological as well as non-vital systems[152] and the distinction it makes is that teleological systems are necessarily directively organized. Thus, teleological explanations are concerned with systems such that their variations satisfy the above assumptions.

The above analysis guarantees now the equivalence between the non-teleological and the teleological explanations that may be given for the evolution of a directively organized and/or goal-directed system. It still seems problematic, though, that in physics there is a preference for non-teleological explanations. The reason for this is that a teleological explanation requires the further assumption that the system under consideration needs to be treated not just as a directively organized system but also as a whole. As Nagel put it [153],

> "teleological explanations focus attention on the culminations and products of specific processes and in particular upon the contributions of various parts of a system to the maintenance of its global properties or modes of behavior. They view the operations of things from the perspective of certain selected 'wholes' or integrated systems to which the things belong; and they are therefore concerned with characteristics of the parts of such wholes, only insofar as those traits of the parts are relevant to the

---

151 Nagel, *The Structure of Science*, p.415.

152 Admittedly, this definition is highly vague and systems, either teleological or nonteleological ones, may be found that do not satisfy the definition. However, the definition "formulates the abstract structure commonly held to be distinctive of 'goal-directed' systems" (p.421).

153 Ibid., pp.421-2.

various complex features or activities assumed to be distinctive to those wholes.

Non-teleological explanations, on the other hand, direct attention to the conditions under which specified processes are initiated or persist, and to the factors upon which their continued manifestations of certain inclusive traits of a system are contingent. They seek to exhibit the integrated behaviors of complex systems as the resultants of more elementary factors often identified as constituent parts of those systems; and they are therefore concerned with traits of the complex wholes almost exclusively to the extent that these traits are dependent on assumed characteristics of the elementary factors.

In brief, the difference between teleological and non-teleological explanations, as has already been suggested, is one of emphasis and perspective in formulation"

Hence in teleological explanations the focus is on the *entirety of a physical structure* and the characteristics of their parts are studied only to the extent that these explain the behavior of the whole. On the other hand, in non-teleological explanations, where we omit the assumption that the physical systems or structures are directively organized and hence we may study the sub-systems of our physical system separately, the focus is on factors that affect *specific parts of a physical structure* rather than the *whole.* To sum up, in both cases we study causal factors and processes, yet in the first case one adopts a *holistic approach* while in the latter a *bit-by-bit* or *fragmented approach.*

An example of such a physical system which can be described only as a (functional) whole is an insulated conductor of an arbitrary shape[154]. When charge is brought to the conductor, it will distribute itself on its surface of the conductor so that the surface forms an equipotential, while at the same time the charge density on the surface is not uniform; it depends on the shape of the conductor. As a matter of fact, the charge will be distributed so that areas with greater curvature have greater density and those with smaller curvature have smaller density. The interesting feature of this system is that the pattern of the charge

154 The example was first given by Köhler (1942) and reproduced in Nagel (1961), p.391.

distribution on the surface of the conductor cannot be built bit-by-bit. In other words, if we brought charge to one part of the conductor and then to another and then to another, thus trying to build the pattern it finally has, we would find that each amount of charge, however small, would distribute itself on the surface so that the density pattern was the one we described. Along the same lines, if we removed some of the existing charge from one part of the surface, the remaining charge would redistribute itself so that the surface would still be an equipotential and the distribution as described.

Other examples of physical systems[155] that behave as a whole are the surfaces assumed by soap films. Given the boundary condition, soap films will form surfaces of minimum area. So, a soap bubble will assume the shape of a sphere as this is the shape with minimum surface for a given volume. If we could remove part of this sphere with circular boundaries, the surface would turn to a plane, as this has the minimum surface for the given boundaries. On the other hand, if we could bring and attach another spherical bubble to the first one, the two would give a new sphere of greater volume. In both examples it is obvious that the conditions (*i*)-(*iv*) hold, so the systems can be considered to be directively organized ones.

It is imperative, for the purposes of this thesis, that we address at once the question whether a constrained Hamiltonian system could be considered as such a directively organized system and this for the reason that we want to classify a specific type of explanation that arises in such systems. Nagel considers the example of a simple pendulum -a bob suspended by a string, experiencing gravitational forces- that is affected by a gust of wind.

---

155 This example is due to Nagel (1961), p.392.

The system S is initially in a state of equilibrium G. The variables we need in order to fully describe the system are the independent coordinates and the forces acting on it. When the wind blows, the bob performs oscillatory motion due to the forces acting on the bob and these are the gravitational attraction, the tension of the string -or the force due to the constraints of the system- the coefficient of dumping and the impulsive force of the gust of the wind. The gravitational force, the damping and the tension result in the so called restoring force. Nagel asserts that this system fails to be a directively organized system because the restoring and the impulsive forces acting on it are not independent, as it was expected in order to satisfy assumption (i) for the variables, because as soon as we know the impulsive force, we also know the restoring force. *But* if we considered that the impulsive force is some environmental causal influence and the restoring force is just the response of the pendulum to the change of its position, then the pendulum can be considered as a directively organized system indeed. This is possible to do if we make the following alterations to Nagel's account.

The system S is a pendulum in equilibrium and the various external forces acting on it; this state of the system is G. Consider that all the causally relevant parts or processes are: the pendulum (along with the forces acting on it when it is in equilibrium) and the environment. The environmental forces acting on it and the 'internal' forces are independent in the sense that we could vary either of the two parts independently. Yet, for whatever primary variation there is an adaptive variation as follows. If we vary one of the environmental forces acting on the pendulum, say if there is a sudden gust, then the remaining forces from the environment (i.e. the resistive forces) along with the 'internal' forces of the

pendulum will vary adaptively, and in accordance with assumptions (*ii*)-(*iv*), so that at a later time the system will be in state G again. Hence, the system is a directively organized system.

The pendulum is only a specific case of a constrained Hamiltonian system. The question we addressed previously, though, is concerned with general constrained Hamiltonian systems: are they directively organized systems too? If we consider that only the unconstrained degrees of freedom are independent, in the sense of assumption (*i*), and in addition to that if we take into account all the laws, principles, environmental factors etc. that constrain the system further, we could claim that such a system is a directively organized one and hence treat it as 'whole'. This way, one may tell a nice causal story and hence give a very good holistic explanation about certain events that occur in a physical structure taking into account what's going on in the entire structure and not just in some small part of it. In the first place, it was not the word 'teleological' that we found most appealing here, rather, it was the word 'holistic'. Yet, Nagel's proposal to understand goal-directed systems as directively organized ones that do not need purposes and goals as dynamic agents may allow us to accommodate explanations from physics that do not fit any of the other suggested models of scientific explanation. Even more, this model may be able to embrace explanations having some of the characteristics of the D-N or the C-R models but not fitting them fully.

4.2 Abstraction, Approximation and Idealization: the Laws of Physics do not Lie, it's Just that the Mappings are Not-All-Inclusive and Non-Exact

We turn now to a very important aspect of explanation, namely that the explanandum is often explained only up to some degree of approximation and correspondingly the explanans may not be strictly true. Although at this stage the link between this section and what follows may not become apparent, after we have discussed the A-B effect and its suggested explanations we will come back to the notions of approximation and idealization and then the hidden link will be revealed. To preempt the reader, though, let us just say that a certain gloss of the topological explanation of the A-B effect will turn out to be non-exact, hence we will criticize it on the basis of what is generally accepted as a fair approximation.

4.2.1 Galileo and the Problem of Accidents

Since Aristotle, who claimed that science's aim is to discover the essences, there seems to have been made a distinction between accidental and essential properties of physical objects. This very distinction is also important for Galileo, although, as Koertge points out, "his conception of accident is interestingly different from Aristotle's"[156]. In this section we will focus on Galileo's views on accidents[157] and on the process that leads from observations of phenomena infected by accidents to discovery of the essences.

Galileo was talking about three different types of accidents. The first is what he called *physical accidents* and these consist of irregularities operating causally in real physical sit-

---

156 N. Koertge, *(1984)*.

157 The reader is referred to Koertge's paper for a detailed analysis.

uations but are deliberately ignored by the theory. One example of a physical accident are the frictional forces acting on an otherwise freely falling body. Then there are the *accidents of observation*, that is to say certain factors involved in the observation that limit the precision of our perception. "Perhaps the most dramatic example", Koertge writes, "is the case of irradiation in which adventitious rays from the stars are refracted by the moisture in our eye and make the stars appear to be twinkling and larger than they really are"[158]. Finally there are the *mathematical accidents*, which are nothing other than discrepancies between the properties of mathematical objects and the properties of the physical ones. So, a real spherical object is not a 'real sphere' whose surface points are all equidistant from its centre. These accidents, according to Galileo, hide and obscure essences and so the naive observer cannot discover them. Throughout his life, then, "Galileo struggled with what [Koertge calls] 'the problem of accidents': because of physical, observational and mathematical accidents we do not find nor expect to find an exact match between ideal, simple scientific laws and what we actually observe. How then can we use experience to appraise our proposed scientific theories?" During this lifelong struggle, Galileo passed through various stages of reflection on the problem which we could roughly summarize as follows. He supported the view that science should be both mathematical and based on experience, yet one should give proofs which are less mathematical and more physical since one would then use assumptions based on observed matters of fact. Whether these assumptions are legitimate, though, depends on our ability to foresee and remove accidents, physical to begin with. Nevertheless, one should not expect theories to match exactly the real world

158 N. Koertge, *(1984)*

experiments because theories are idealizations, so there will always be an observation gap caused by physical accidents. Laws, which are the result of such idealizations, hold only for accident-free situations. Moreover, the gap widens by the presence of mathematical accidents or mathematical approximations that, once again, we inevitably make. One way for removing accidents, he suggests, is by improving our experimental techniques, whenever this is possible. When this is not possible, like in the case of the omnipresent frictional forces for example, one may vary the 'degree' of accident and check the results. When the accidents are small and irregular, one has to just ignore them. And he goes on to suggest that in certain cases we may even have to abstract from major interferences, almost as big as the effect itself. Two things would make an answer probable to Galileo: simplicity considerations and, most significantly, whether the theorems on which the answers were based were anchored on observation and experimentation.

So, we could summarize Galileo's beliefs about how one may arrive at a theory as follows. Since one has to deal with accidents, of which one may find an infinite amount, it is necessary to abstract from them and then use the abstractions with the limitations that experience teaches us. In order to abstract, the 'recipe' to follow is the this: vary the degree of perturbation, note the resultant effect, extrapolate to the limiting case where perturbation is absent. In this process approximations are perfectly legitimate whenever the effects are too small -at least comparatively. Also, one has to deal with experimental error by doing controlled experiments and by eliminating as much as possible the accidents of observation.

An idealization, which is the result of this process, is the "quantitative extrapolation from real-life experimental situations to the ideal-theoretical-limiting case"[159].

What one can undoubtedly notice here is that for Galileo, there is a two-way process going on when doing science. One first goes from observation and experimentation to idealization, through abstraction (or elimination of accidents) and approximation. Then, starting from theory, one may go back to observation using approximations whenever we cannot improve the experimental situation any further, and acknowledging and accepting that the experimental results will never match exactly the theoretical predictions. Galileo's views about how one arrives at a theory -an idealized view of the world?- is similar to Shapiro's views about how we ever get to know abstract mathematical structures. If Shapiro gets it right about mathematics, as we believe he does, and if Galileo gets it right about physics, then new theories in both disciplines are likely to be inspired by observation of the same physical systems from which they abstract. Hence it should come as no surprise that often, and even in cases where there seems to be no dynamic interaction between physics and mathematics, it is as though mathematical theories, already developed, have been waiting on a display for several years before they are picked to be employed in the formulation of some physical theory. After all, a deep connection between mathematics and physics can be found in the very ideas or intuitions that lie in the very foundations of the theories and in the fact that abstraction follows similar paths in both.

Before we conclude this section we would like to make two further remarks. One concerns the surplus structure that certain physical theories acquire once they are expressed

---

159 Ibid.

in mathematical language. Apart from the kind of abstraction that Galileo is talking about, which results in an 'impoverished' version of physical systems, there seems to be some different, albeit parallel, process that results in something reminiscent of what we have called surplus structure. Koperski in his (2001)[160] suggests that artifacts -which correspond to objects belonging to our surplus structure- "are the false properties or relations that can result from idealizations. An artefact is not an abstraction built into the model; it is a (possible) consequence of simplifying assumptions"[161]. Consequence of the simplifying assumptions, or of something different, the point is that those objects often appear to have an explanatory and predictive power that we would like to try and explain in the last sections of this chapter. Partly, the power of the surplus structure may be justified if we considered it to be necessary for encoding all the information needed for the description of the physical system, without corresponding to some real physical entity.

4.2.2 Models and Analogies in Science

Mary Hesse, in her Models and Analogies, claims that there is more to models than being just mere aids to theory construction, as a Duhemist would suggest. Adopting a Campbellian view, she asserts that theories are expected to fulfill more than just being a mathematical system with deductive structure. A theory, if it is to be an explanation of phenomena, ought to be intellectually satisfying in the sense that it provides interpretation in terms of models, to be mathematically intelligible and maybe simple and 'economic'. Moreover, theories are dynamic in the sense that they are extended and modified in order to make pre-

160 In this, Koperski follows Wilson (1991).

161 J. Koperski, 2001.

dictions and account for new phenomena. According to Hesse, that would not be possible if for extending the theories one did not use analogies with the already existing models, for without models theories cannot be genuinely predictive. Hence models are essential to the logic of science.

A model is analogous to the physical system it models, or to some other model, or to a theory, in three ways, Hesse claims. First, there is the so-called *positive analogy* which refers to the properties of the model that are found in the system as well, then there is the *negative analogy*, which reflects properties of the model that are not found in the system and finally there is the *neutral analogy,* which is what allows for predictions and which refers to analogies that we do not know yet whether they are positive or negative. She emphasizes that while in an accepted theory we will find only positive analogies between theory and a physical system, in what she calls $model_1$, which "*is* the way we are imagining the phenomena themselves"[162], we will find both positive and neutral analogies, whereas in the so-called $model_2$ we may find all the three types of analogy present. The observed properties and the observed analogies between models and physical systems, or models of one physical system and models of another physical system, are the sources of information that help both in explanation and in theory construction. But the explanatory role is played by the positive and the neutral analogies only; in her own words, "[w]hen we consider a theory based on a model as an explanation for a set of phenomena, we are considering the positive and neutral analogies, not the negative analogy"[163]. In addition, it is the neutral analogies that have predictive strength and hence may show the way towards new theories.

---

162 M. Hesse, *Models and Analogies*, p.11.

163 Ibid.

These analogies between the properties of two analogues we may call '*horizontal*'. Analogies have one more feature: there are relations between the properties of the same object or model that are causally linked, which we will call '*vertical*'.

Hesse distinguishes further between two types of analogy which she calls *formal* and *material*. A formal analogy is a one-to-one correspondence between different interpretations (or models) of the same theory. It is a post-theoretic analogy in the sense that it can be identified as such after the theory has been established and the models have been invented. On the other hand, material analogies are pre-theoretic analogies between observables; pre-theoretic in the sense that such analogies can be identified between two models before a theory has been established for one of the two, the one that we call the explanandum. Material analogies between established model and explanandum, then, enable scientists to make predictions of a new theory.

Let us examine now how material analogies combine with positive and neutral in explanation and how they are used in predictions. To do that, we will use the following example. Suppose that we are aware of the wave theory which is expressed by the wave equation $y = a \sin 2\pi f x$, where a is the amplitude of the wave and f its frequency. We are also aware that the theory is interpreted successfully in what we could call sound model$_2$, which contains all observational properties, such as loudness, pitch, detected by ear, propagated in air and so on. Furthermore, we acknowledge that light observables like intensity, color, propagated in aether and so on may be interpreted using the same wave theory. Hence, before we establish any theory of light, we recognize that there exist pre-theoretic material analogies between model$_2$ and light observable properties. Rendering this bunch

of properties of the sound waves into a $model_1$, we may now attempt an explanation of the wave properties of light in terms of a new wave-theory of light that will be based on the positive and neutral analogies between $model_1$ and $model_2$. This theory leaves out of the analogy strategy the negative analogies that inevitably exist between the sound-$model_2$ and the light-$model_1$. The fact that there are some negative analogies present in $model_2$ is not sufficient by itself to ban the model, so far as the properties to which they refer and which they causally affect are not essential[164]. What is important to point out, though, is that in predicting and probing the wave theory of sound to account for light, one begins of course with the known positive analogies, but one has to rely on the neutral analogies to formulate hypotheses that may or may not be refuted afterwards by experimental evidence. So, the 'horizontal' similarity relations that hold between the two models allow for predictions and inferences in the 'vertical', causal direction; predictions that involve the neutral and maybe the negative analogue properties of $model_1$.

4.2.3 The Chaos Case

In *Explaining Chaos*, Peter Smith addresses the question of what constitutes a good scientific explanation of chaotic systems and his case study is very relevant to ours in that in chaos theory, the mathematical structure involved has got what he calls *surplus content*, which is very similar to our own surplus structure. In his own words, the mathemati-

[164] Hesse does not provide a clear-cut answer to what it means for a property to be essential. But she does consider the following three suggestions (p.100-1). First, essential are properties that are causally closely related to the positive analogy of the model. Second, if a property is so closely related to the neutral analogy that it would render it negative if the property in question was shown to be negative, then it is also essential. Finally, a model with some negative analogy may be retained even when the negative analogy affects the neutral, just in case there are no alternative models available.

cal structure with the surplus content "is like a map with an unlimited amount of excess, necessarily fictional, content"[165]. The problem that one faces is that the theory, which he considers to be an idealization of nature, provides models which "misrepresent the facts by involving patterns of dynamical behavior which have an intricacy that the modelled phenomena must typically lack"[166]. How, then, could such a theory provide explanations? His response comes in three parts.

First of all, he considers that chaotic theories can be "richly predictive in a variety of ways", hence useful in the sense that they put the theory back into empirical work and for that reason they may reveal correlations between parameters and dynamical features that do play an explanatory role. Moreover, though not strictly true, they are still approximately true. One reason why a theory, in general, and chaotic theories, in particular, cannot be strictly true is that the infinite theoretical precision of the idealized theory will be only and always met by finite physical/experimental accuracy, he claims, and with this claim he reminds us of Galileo's mathematical and observational accidents. In order to define the notion of approximate truth, Smith distinguishes between two different types of structure. The one, which represents, consists of the bundle of abstract trajectories and it is the structure that is doing the actual modeling, while the other, which is represented, is the structure encoding what needs to be modelled and it consists of all the physically possible time evolutions of the real-world dynamical system. "If these two are replicas, then we say that the dynamical theory that postulates such a model is true, period. And if the structures are similar enough, we can say that the dynamical theory in question is approximately

165 P. Smith, *Explaining Chaos*, p.43.

166 Ibid., p.51.

true"[167]. Of course this rough definition and the phrase 'similar enough' in it opens up a whole philosophical debate[168], but this is beyond the scope of this thesis and hence we will not pursue it. For his own purposes, the definition does the job of attributing to the models approximate truth. Finally, he points out that certain properties -like for example period-doubling leading to apparent chaos- present in the mathematical model are *universal* in the sense that they are shared by a wide class of physical cases as well, which are empirically observed, and this is, of course, reminiscent of Hesse's analogies. So what requires an explanation "is how the universal features of a family of discrete maps are related to the modelling of real-world continuous processes"[169]. Examining the models with the universal feature and the relation of the parameters responsible for the resulting chaotic behavior to the theory as a whole, Smith concludes that we are getting a partial explanation of why the dynamics turns out to be chaotic by referring to more general principles of the theory. One objection to this partial explanation might be that it is qualitative rather than quantitative, but this is not alien to scientific practice, he claims.

An issue that is related to both the notion of approximate truth and the notion of universality is that according to a certain equation of the theory, the Navier-Stokes equation, very small changes in the initial conditions can have unpredictably big effects. The models that are based on this equation are usually derived by throwing away all the higher order terms that are responsible for the unpredictably big effects. But then the problem one faces is one of credibility of the resulting approximation. Smith's response is that "some

---

167 Ibid., p.72.

168 For detailed discussions see P. Smith, *Explaining Chaos*, D. Lewis, *On the Plurality of Worlds*, D. Miller, *Critical Rationalism*.

169 P. Smith, *Explaining Chaos*, p.102.

features of the model may be relatively *robust*, i.e. be features which are also shared by variant models where other perturbing terms are thrown in to make the defining equations somewhat more realistic. And we might be able to appeal to those more robust features to extract useful predictions about the kinds of behavior and the kinds of transition to be found in the physical system. Universality results establish that certain features can be particularly robust"[170]. He suggests, therefore, that the features to be taken seriously are the robust features and not the ones that belong to the surplus content of the theory. Comparing, once again, with Hesse's terminology, we could say that the robust features correspond to positive analogies.

The question that is raised from all that is whether the part of the mathematical structure in chaotic models that does the explaining makes use of a neutral or a negative analogy. And as it is the case in the A-B effect as well, as we shall see shortly, there is indeed a negative analogy at the heart of this explanation, since the mathematical structure that models the physical system, namely the fractal structure, is infinitely intricate, while the physical system is not, apparently. Smith claims that fractal attractors -the negative analogy- do not have to be interpreted realistically and they may even be left out as uninterpreted mathematical objects. This, in Hesse's terminology, means that the positive and the neutral analogies of our model will not be 'causally' affected by the non-inclusion of the fractal attractors and hence the model will not be fatally affected.

Along the same lines with Smith's views are Orly Shenker's who argues in her (1994) "that fractal geometry can only be approximately applied to natural forms" because even

---

170 Ibid., p.125.

when the geometrical structures seem to match physical forms amazingly, these actual geometrical structures are not fractals. Fractals are primitive geometrical objects that possess infinitely many details in a finite volume -i.e. they have infinite complexity- and they may be described as geometrical processes that continue *ad infinitum*. However, she argues, the geometrical objects that are used as representations of physical forms neither possess infinite details nor could the process of their construction continue for ever -there is a cut off that renders them into an approximation of fractals that, in turn, approximate natural forms. Hence, she concludes, "they resemble natural forms due to their *not* being fractals". But even if we accepted as legitimate the approximation of physical forms by an approximation of the actual mathematical structures, the latter have hardly any of the properties of the former. For if one tried to, say, magnify the 'fractals' that resemble a landscape one would realize that the previously apparent resemblance between the two now vanishes. Moreover, chaos theory has no coherent interpretation, hence, no far-reaching physical conclusions can be drawn, and being a theory of infinite detail it is not consistent with the atomic hypothesis. All the above considerations, as examples of what Hesse would call negative analogies, lead her to the conclusion that fractal geometry is not the geometry of nature. To anticipate our discussion of analogies in the case of the A-B effect we would like to point out that despite the similarities that we may find between the two cases, there is a major difference, namely that in the A-B case there is a whole theory related to it and a very successful one indeed.

4.3 Three Attempts for an Explanation of the A-B Effect

The Aharonov-Bohm effect, also known as the A-B effect, is an effect one finds in every quantum field theory book since everybody appeals to it in order to justify why one should consider that the gauge field in electromagnetism is actually a real physical entity. The prediction and subsequently the experimental verification of the A-B effect have been crucial cornerstones in the history of physics since they suggested that the connection, or the A_μ field, might be interpreted as a real field[171], rather than just a mathematical artefact. Hence, ever since its discovery, physicists take it for granted that A_μ does represent something as tangible as -at least- any matter field. But this is just the physicists' view, which means that there is quite a lot of dirt left under the carpet, dirt that we aim at clearing up in this section. But first things first, we give an account of the effect itself, before we attempt to give some explanation for it.

4.3.1 The Effect

The setting for the A-B effect is very similar to the two-slit experiment with just one difference: right outside the two slits and in between them there is a very fine and long solenoid, ideally infinitely long, producing a magnetic field that is confined entirely within the tube of the solenoid.

The configuration in the two-slit experiment is depicted below.

[171] This is what I call the second approach to the A-B effect. According to this approach, the effect may be accounted for by considering that the A_μ field is a real physical field which acts on the passing electrons and causes a phase shift on them.

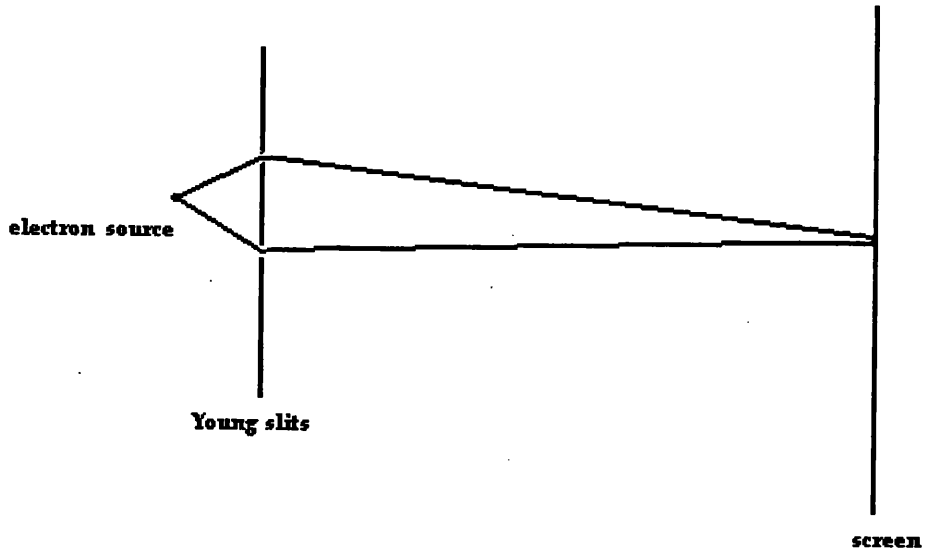


Figure 10
The Two-Slit Experiment

If we consider that electrons may pass from either slit, but each electron passes from one slit only, the interference pattern that appears on the 'screen' of the two-slit experiment may be explained as a result of the phase difference between the wavefunctions of the electrons that arrive there. So, if the phase factor of an electron which passes from slit 1 is $e^{i\Phi_1}$ and the phase factor of an electron that passes from slit 2 is $e^{i\Phi_2}$, then the phase difference of the two 'waves' is given by

$$\delta = \frac{2\pi\alpha}{\lambda} = \frac{2\pi\varkappa d}{L\lambda}$$

where α is the difference in the path length for the electrons going through the two slits, d is the distance between the two slits, $\varkappa$ is the distance from the axis of symmetry of the 'screen', λ is the wavelength and L is the distance between two-slits and 'screen'.

When a solenoid is inserted in-between the slits, the configuration changes as follows.

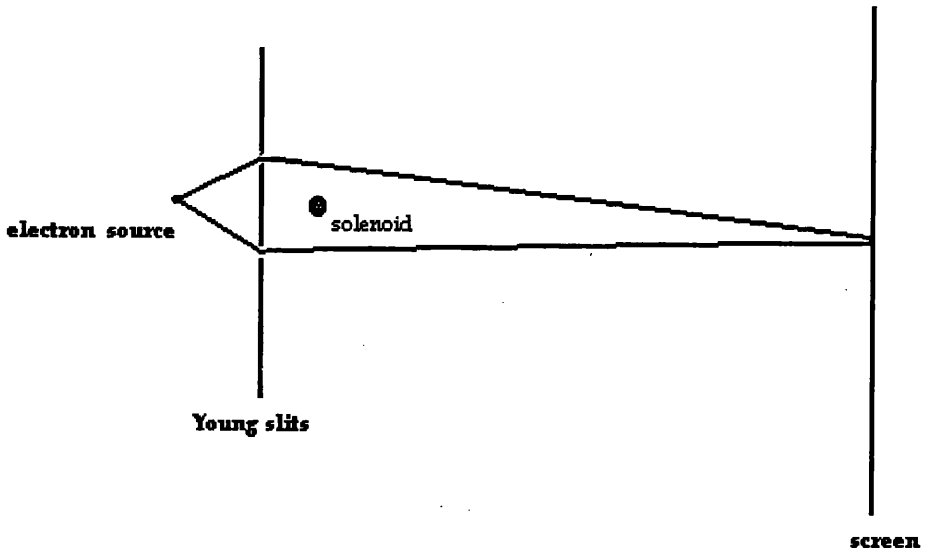


Figure 11
The Aharanov-Bohm Experiment

Now that we have a magnetic field present in the area, there is a change in the phase of the electrons that pass by, which is equal to $\frac{q}{\hbar}\int_{trajectory}\mathbf{A}\cdot d\ell$. So both phases Φ_1 and Φ_2 of the electrons that pass through each of the two slits respectively will become:

$$\Phi_1 = \Phi_1(B=0) + \frac{q}{\hbar}\int_{(1)}\mathbf{A}\cdot d\ell$$

and

$$\Phi_2 = \Phi_2(B=0) + \frac{q}{\hbar}\int_{(2)}\mathbf{A}\cdot d\ell$$

Since the interference of the waves on the 'screen' depends on their phase difference, there will be a new pattern determined by the difference of the new phases:

$$\delta = \Phi_1 - \Phi_2 = \delta(B=0) + \frac{q}{\hbar}\int_{(1)}\mathbf{A}\cdot d\ell - \frac{q}{\hbar}\int_{(2)}\mathbf{A}\cdot d\ell = \delta(B=0) + \frac{q}{\hbar}\oint\mathbf{A}\cdot d\ell$$

This equation tells us how the electron motion changes when some magnetic field is present. (NB. *Any choice of A which has the correct curl gives the correct physics.*) If we use Stoke's theorem at this point, the equation above becomes:

$$\begin{aligned}\delta &= \Phi_1 - \Phi_2 = \delta(\mathbf{B}=0) + \frac{q}{\hbar}\oint_{\partial s}\mathbf{A}\cdot d\ell = \delta(\mathbf{B}=0) + \frac{q}{\hbar}\int \nabla\times\mathbf{A}\cdot ds = \\ &= \delta(\mathbf{B}=0) + \frac{q}{\hbar}[\textit{flux of B between (1)and(2)}]\end{aligned}$$

Since the flux of **B** does not depend on which pair of paths we may choose, provided that they surround the solenoid, for every arrival point there is the same phase change $x_0 = \frac{q}{\hbar}[\textit{flux of B between (1) and (2)}]$. This means that the entire pattern is shifted by x_0.

These are the general ideas involved in the A-B effect and the main consequences are two. The first one is that if the magnetic field, which is confined inside the solenoid only, accounted for the effect, then we would have action at a distance that, clearly, violates locality. The second is that the **A** field may thus be considered as real in the sense that "it is what must be specified *at the position* of the particle(electron) in order to get the motion"[172].

Since the Aharonov-Bohm original publication in 1959 quite a lot of discussion has been going on about the effect itself and its experimental verification. Some, like for example Bocchieri and Loinger (1978), have challenged the validity of at least the early experiments and even have gone as far as to claim that the effect does not exist. The early experiments, conducted in the early 1960s, made use of very thin solenoids and whiskers[173], but their validity was challenged on the basis that since the solenoids were not infinitely long, there should be some 'flux leakage' from the two ends, which in turn would affect the

---

[172] *Feynman Lectures on Physics*, II-15-12.

[173] Whiskers are very fine permanent magnets with diameter of the order of $1\mu m$. The magnetic flux inside the whisker is proportional to its cross section. The idea in these experiments was to put a tapered whisker in the shadow of a solenoid and check the deflection of the electrons.

electrons and could be held responsible for the effect. In response to criticisms and in order to avoid possible leakages, toroidal solenoids were used in later experiments, and their use enabled the experimenters to measure the effect of the potential -rather than the effect of the field that leaked- with undisputable accuracy[174].

4.3.2 The Three Attitudes Towards the A-B Effect

There are three (plus one) different ways of explaining what is going on in the A-B effect, but they all meet philosophical reservations and criticism. The discussion and philosophical examination of the fourth one is our contribution to the debate[175]. A common physicists' story, which is a paraphrase of the conclusion of the previous section, contains two out of the three approaches, and goes as follows. Given the facts, there are two possible ways of explaining what is going on. According to the first one, if we take the magnetic field **B** as the only existing interactive field, then we would have to succumb to action at a distance and hence to non-locality. To remedy this action at a distance thing, which no one really likes because one cannot tell a nice causal story that explains the facts, *one has to consider as true* the assumption that the physically interacting field is the vector potential **A**, and this constitutes the main assumption of the second approach. At first sight, this second approach is problematic because the **A** field which is held responsible for the effect is not gauge invariant and hence if we were to consider as real only the gauge invariant objects of the theory, this one does not qualify. In the third approach, one considers as the real causal

[174] For a detailed discussion of the A-B effect, the experiments conducted to measure it and the discussions that followed them see Peshkin & Tonomura, *The Aharonov-Bohm Effect.*

[175] We would like to stress here that the fourth way has been discussed in the physics literature but not in the philosophical. The discussion of the fourth way from a philosophical perspective constitutes our own contribution to the debate.

agent the so called Dirac phase. But taking a slightly closer look at these arguments we find that there is more that needs to be said in order to make this a good explanation of the effect and quite a lot that is missing.

The Three Ways as Discussed in Healey and in Lyre

All the three approaches are discussed in some detail in Healey (1997, 1999, 2001) and in Lyre (2001).

Healey claims that since the A-B effect involves some kind of interaction between either electromagnetic fields or potentials and electrons, if *either* the interaction *or* the fields *or* the potentials are not local, then neither is the effect. His aim, then, is to show that *in both* the cases we mentioned above there is violation of locality of some sort or another[176]. But before we go on to examine Healey's argument, we need to make clear what he refers to when he talks about locality.

Locality and Separability in Healey

Healey, in accordance with Einstein, discerns two different notions concerned with locality, both necessary for a process to be local. He calls them *local action* and *separability* and he gives them the status of principles. So, for him, locality holds just in case both local action and separability hold.

---

[176] Note in passing that in his (1997) paper Healey not only investigates the notion of locality in the quantum domain of gauge theories, but he also compares and tries to draw the parallels between the A-B effect and the case of the Bell inequalities. The focus of this thesis is on the former aspect so we will not refer at all to the comparative aspect of Healey's work. However, for more information see Tim Maudlin (1998).

The principle of local action is expressed as follows: "If A and B are spatially distant objects, then an external influence on A has no immediate effect on B"[177]. A little later, Healey writes that "the idea behind local action is that if an external influence on A is to have any effect on B, that effect must propagate from A to B via some continuous physical process. Any such mediation must occur via some (invariantly) temporally ordered and continuous sequence of stages of this process"[178].

What Healey means when he talks about 'the principle of local action' is that if A and B are things separated in a way that they cannot influence each-other instantaneously, there must be some physical process that propagates the effect of some influence from A to B. This process can propagate with some finite velocity (less than or equal to the velocity of light) and, therefore, it can influence B only after the lapse of some finite time interval. In relativistic language, this leaves the influence within the light cone and maintains the causal order for observers in all inertial frames. So, in order for local action to hold, two requirements must hold: *influences (i) are mediated by physical processes and (ii) propagate with sub-luminal velocities.* These two necessary conditions re-express the principle.

Violation of Local Action in the First and the Second Ways

Let us assume, first, that electromagnetism is described by the 'real' (electro)magnetic fields, in accordance with the first way of understanding the effect. In this case, the principle of local action entails that a change in current (in the solenoid) has immediate (i.e. not

177 Healey, 1997.

178 Ibid.

mediated) effect on the electrons outside it, because as we have seen, the magnetic field is confined within the tube of the solenoid, and this means that this field itself cannot 'mediate' an influence affected by the change of the current and hence of the magnetic field. Since the only physical entity we are considering here is the magnetic field, the influence is not mediated by a physical process and, therefore, it violates the principle of local action.

On the other hand, if we assume that the A-B effect is accounted for by the gauge covariant vector potential, we face two difficulties. The first one is that the **A** field is not a physically real field, so how it could mediate anything at all, and the second is that, regardless of its physically non-real nature, it does not act on the electrons directly either! But more on the violation of local action by the **A** field later.

Separability

A common understanding of separability involves 'entangled' quantum systems, which are non-separable in the sense that they must be described by a tensor-product state-vector which does not factorize into a vector for each of the individual systems that compose it. i.e. $\Psi_{12...n} \neq \Psi_1 \otimes \Psi_2 \otimes ... \otimes \Psi_n$. The non-factorizablity, on the other hand, means that the state of the system whose constituents are the $\Psi_1, \Psi_2, ..., \Psi_n$, does not supervene on the states of its constituents; in other words, knowing the states of the constituent systems does not suffice to know the state of the entangled system[179]. Hence, in this common understanding, two or more, spatially separated systems are non-separable *if and only if* the state of the compound system does not supervene on the state of each of the constituents.

[179] In fact, the constituent states in a composite entangled state are not even pure states but the so-called improper mixtures.

Nevertheless, Healey is up to some different notion of separability[180], which is based on what he calls the 'principle of separability' and it does not refer to entangled quantum systems only. The principle is expressed as follows: "Any physical process occurring in spacetime region R is separable just in case it is supervenient upon an assignment of qualitative intrinsic physical properties at spacetime points in R".

Of course the notions 'qualitative' and 'intrinsic' are far from being straightforward, and Healey is well aware of this fact. He suggests, though, an intuitive and inconclusive resolution. *Intrinsic*, he says, is a property that an object has in and of itself. For example, the presence of some specific magnetic field both inside and outside the core of an electromagnet is an intrinsic property of the electromagnet. (compare extrinsic, properties that an object has in virtue of its relations. E.g. the attraction of iron fillings by an electromagnet is not an intrinsic property of the iron fillings, because it depends on the presence of the magnet close enough to the iron fillings.) *Qualitative*, as opposed to individual, is a property just in case it does not depend on the existence of any particular individual. E.g. behaving like an electromagnet does not depend on any particular electromagnet.

Despite the fact that his resolution is not conclusive, qualitative intrinsic properties (QIPs for the sake of brevity) are exactly what science is looking for, he claims. Then science characterizes the various objects as certain kinds of physical systems and specifies their state by ascribing to them those properties. Fundamental physics, in particular, which investigates the basic kinds of physical systems, aims at characterizing their states completely, so that the physical properties of the more complicated systems that these constitute

[180] For a detailed discussion for the difference between the two notions of separability see Healey (1997, 1999) and Maudlin (1998).

are then determined. In other words, Healey believes that the properties of the complicated systems supervene upon those of their more basic constituents, and it is in this sense that systems may be separable, that is to say, *just in case* their properties supervene upon those of their constituents. Physical processes, on the other hand, consist of suitably continuous sets of stages that involve one or more enduring (physical) systems. Thus, the physical processes are separable *just in case* they supervene upon qualitative intrinsic properties of (objects) at spacetime points in the region where they take place.

Violation of the Principle of Separability in the Second and the Third Ways

The (electro)magnetic field in the A-B effect is non-local in the sense that it violates the principle of local action. Yet, if we adopted the second view, namely that the interactive field is the potential, in order to settle the locality issue, the explanation would have to meet the challenges that the above notion of separability has in store. And this means that if either the process by which each electron passes through the region outside the solenoid or the electromagnetic potential there throughout the time of its passage do not supervene on QIPs of (objects at) points in that spacetime region, then the alleged local explanation of the effect violates the notion of separability. For this reason, he examines how separability is challenged by some 'acceptable' form(s) of the gauge potential, first, and then by the process by which the electrons pass through the apparatus. A very similar approach to these two notions we find in Lyre's approach as we shall see shortly.

Healey's argument is the following. The A-B effect involves some kind of interaction between electromagnetic fields or potentials and electrons. If *either* the fields *or* the

potentials *or* any other mathematical quantity we use for the explanation of the effect are non-local, then so is the effect[181]. *Local action is violated both by the electromagnetic field and by the A field itself if we take it to be real. Separability, on the other hand, is violated by the A field as well as by some other, gauge invariant, form of the A field. Therefore, in either case, the two approaches are characterized by non-locality.* And here is how he supports the above conclusions.

First, he shows that if we take the magnetic field $\mathbf{B}$ to account for the A-B effect, then we have obvious violation of local action, without much ado. The magnetic field is confined inside the solenoid while the electrons pass outside it. So this non-locality is just a straightforward consequence of the electromagnetic theory and the particular experimental set up, as we have already seen.

Then, he goes on to show that the 'bare' $\mathbf{A}$ field does not act on the electrons locally either. The electrons, he argues, follow specific paths. The shift of the interference pattern in the A-B effect is produced by a direct local interaction between electrons and the gauge potential outside the solenoid. The A-B effect is local *only if* $\mathbf{A}$ is a physically real field *and* it is capable of acting on the electrons directly. But since A_μ is a gauge dependent field, it is not a physically real field, because the physically meaningful quantities must be gauge invariant. $\mathbf{A}$ is not gauge invariant, which means that both $\mathbf{A}$ and $\mathbf{A}' = \mathbf{A} + \nabla x$ (*should*) specify the same physical state. Hence the A_μ field is not a physical object. As Maudlin (1998) pointed out in his response to Healey's (1997) paper, the soundness of this explanation of the effect depends on his interpretation of gauge theories. This is quite an

[181] His initial claim is that if the effect is local then either the E&Ms or the A's or the process are local.$(C \rightarrow A_1 \vee A_2 \vee B)$. So, $\neg(A_1 \vee A_2 \vee B) \rightarrow \neg C)$.

important point but we will come back to it shortly. For the time being, let us carry on with Healey's argument.

"If one nevertheless maintains that in some way **A** represents a physically real field", he continues, "the following argument appears to establish that its gauge-dependence excludes local action"[182]. Assume that, somehow, **A** is a physical field capable of carrying the influence from a change of the magnetic field inside the solenoid to the electrons that pass around it. But **A** does not act on each electron directly, because each time an electron follows a particular path we can choose a local gauge transformation that sets the gauge zero along that path. Obviously, this approach violates the principle of separability and hence the description is still non-local.

Maudlin, in his (1998), discussing precisely this point claims that Healy's "argument establishes nothing at all" because in theories where the wavefunction is complete, the electrons take both paths around the solenoid and even if one considered theories where the electrons take specific paths, the electron-wavefunction is still affected and hence influences the path. But although local action may thus be established, still, the physical reality of the gauge fields is not established because they are not gauge invariant quantities. Maudlin suggests that gauge freedom, along with the question "why gauge invariance is a sine qua non for physical reality?" is at the heart of the problem. He then proposes that if one was willing to accept that there is ONE TRUE GAUGE describing the effect at any time, one would have an explanation both local and separable, albeit one would face epistemological inaccessibility -cannot know which gauge by observation- and indeterminism

182 Healey, 1997.

-if the local gauge transformations were considered in an active way. But as we shall see shortly, Gribov (1978) and Singer (1978) showed that even the idea that there might exist a one true gauge is not feasible.

The next step Healey takes is to argue that since the A_μ field does not manage to account for some kind of local interaction directly, and since this happens because of its gauge-dependence, one could expect that some gauge invariant quantity involving A_μ might do the trick. The Dirac phase factor is a good candidate because, after all, this is what measures the phase shift. The Dirac phase factor is expressed by the integral $S(C) = \exp[-(ie/\hbar) \oint_c A(r) \cdot dr]$, where the integral is taken over each closed loop C in spacetime. Hence, Healey considers the integral $I(C) = \oint_c A(r) \cdot dr$ as the quantity that expresses an intrinsic property of C, provided that C is a non-intersecting closed curve[183]. But the problem in this approach is that the $I(C)$s "do not supervene on any assignment of qualitative intrinsic properties **at** spacetime points **in** the region concerned", because by its definition each $I(C)$ supervenes upon the spacetime points of an arbitrary curve $C = \partial s$ which encircles the solenoid and not on the spacetime points through which a single electron passes. Therefore, he concludes, if we choose the loop integral $I(C)$ to describe the A-B effect, we have violation of separability because for a physical process to be separable, it must supervene upon an assignment of qualitative intrinsic physical properties at spacetime points that define the trajectory of the electron. So, "irrespective of the quantum description of the electrons, the A-B effect manifests non-locality either because it is

[183] He takes$I(C)$, rather than $S(C)$, in order to get rid of the electronic charge e, and he chooses non-intersecting closed curves in order to avoid the difficulty arising by the fact that closed curves do not correspond uniquely to regions of space.

taken to be completely described by the electromagnetic field (i.e. violation of local action), or because electromagnetism is taken to be completely described by (something like) the Dirac phase factor (i.e. violation of separability)"[184].

Comments on Lyre's Approach and Beyond

Lyre's approach is very similar to that of Healey. In his paper *A versus B! Topological Non-Separability and the A-B Effect (2001a)* he too talks about the same three approaches which he calls B, A and C respectively. Using similar notions for local action and separability he concludes as well that the B approach violates local action, while the A and C approaches violate separability; the violation arises because "the observable effect of the shift of the interference fringe cannot be reduced to properties associated to spacetime regions". He claims that this lack of consensus about which explanation is the best -if one exists at all- along with the fact that each one of them has elements not present in the other two are evidence that the A-B effect and its tentative explanation are a typical case of unerdetermination of theory by evidence.

The 'loopy' or C approach, which is favoured by Lyre, as well as by Healey in his most recent work (2001), is based on precisely the realization that the A-B effect is a *'global'* effect and to our view this is a good attempt to take the global nature of the phenomenon into account. We put the word global in inverted commas because it is a little bit too heavy for the actual meaning it has in this context. By that we mean that the word global in this context means comprehensive or inclusive, *and not* universal, in the sense that the net effect on the phase of the electron is the result of the loop integral $I(C) = \oint_c \mathbf{A}(r) \cdot d\mathbf{r}$

[184] Healey, 1997.

along a curve that surrounds the solenoid, which is also known as the *holonomy*. The curve along which we integrate is arbitrary and can get as close to or as far from the solenoid as we like; it is in this sense that the phenomenon is global, and not universal. Lyre, along with Drienschner and Eynck in their (2001) define *prepotentials* to be "non-separable equivalence classes of gauge potentials in the whole of space" and consider the prepotentials to be real on the basis that if they are altered there is a physical effect and that they act locally, since they are to be found where the electrons pass from, though non-separably. To their view, prepotentials provide the proper ontological description of the A-B effect when they are considered to be the fundamental entities in gauge physics and their use has the advantage that "avoids the introduction of mysterious surplus structure"[185].

Although this is a good attempt for an explanation of the A-B effect, there are a few misunderstandings in it, we believe. First of all, the prepotentials as defined in Drienschner, Eynck and Lyre's paper are not exactly the same object as the holonomies, even though the two are related. The gauge fields are, as we have seen, the Lie algebra valued one-forms, while the holonomies are their loopy integrals $I(C)$. Using Stoke's theorem for the electromagnetic case, we see that

$$I(C) = \oint_c \mathbf{A}(r) \cdot d\mathbf{r} = \int \boldsymbol{\nabla} \times \mathbf{A} \cdot d\mathbf{s} = \int \mathbf{B} \cdot d\mathbf{s} = [\textit{flux of } \boldsymbol{B}],$$

or in words, we see that the holonomy, i.e. the phase-integral around the loop is the same quantity as the flux of the magnetic -or curvature two-form- field from a surface that intersects the solenoid and whose boundaries surround it; this is nothing other than the horizontal lift of the wavefunction when parallel transported over a closed curve. Of course

185 Lyre, (2001a).

one could claim that a prepotential, that is a specific equivalence class of gauge potentials, is real in the sense that Lyre attributes to the term real, but then this only says that if we alter the class, we describe a different magnetic field which will have a different physical effect, of course. From this perspective, therefore, the prepotentials contain exactly the same amount of physical information as the magnetic field itself. If Lyre considers the prepotentials to be identical with the holonomies, on the other hand, then in them there is something more, namely quantitative information about the horizontal shift as we have said. However, even that does not give a good reason why holonomies could be considered as physical objects for they only measure a shift, after all. It is hard to see how something that is not physically detectable, something that constitutes just a measure of the effects of the parallel transport of a physical object along a closed curve, may be given the status of a physical object. At the same time, as mathematical objects they signify properties of space-time that, in turn, describe or even determine, one might say, the effects of some sources on the electrons that pass around them. Finally, we would like to remark that the use of holonomies does not avoid surplus structure, for holonomies themselves do come in equivalence classes of mathematical objects that do not correspond directly to physical objects and define a transformation group, the so-called holonomy group. The very occurrence of a transformation group signals the presence of ambiguity of representation of either second or third type, and since the transformations in this case cannot be taken to be active, it is definitely of the third type, hence there is surplus structure involved.

So, what the above discussion leaves us with is that the A-B effect is inherently non-local and this is a characteristic that any good explanation of it needs to account for.

Although it doesn't follow that *we have to* adopt a holistic explanation, attempting to give one is a good bet because we need to explain a global effect. Lyre et Al, write in their (2001) that "this indicates the deep topological nature of the A-B effect -stemming from the topology of the gauge group $U(1)$", while Lyre himself writes in his (2001a) "were it not for the non-trivial topology of *both* the base space and the gauge group, any two magnetic fields confined to the inside of a solenoid would necessarily have to have the same (null) effect on the interference pattern. Therefore, only the non-trivial topology of both spaces produces the A-B effect and its peculiar type of nonlocality is best addressed as topological non-separability". With these comments, Lyre et Al rather confuse the holonomy approach, which does not involve topological considerations in explaining the phenomenon, with our fourth way, which is a purely topological interpretation of the effect. However, they anticipate the fourth way and indicate that the holonomies are linked to topological considerations that, we will argue, justify their usage in an explanation of the effect. This justification will become clear, we believe, once we have discussed holonomies from the perspective of fibre bundles, a discussion that will illuminate two things: first, the fact that holonomies describe a change rather than producing it and second, the relation between holonomies and one attempt to provide a topological explanation.

Although we are already able to see how the need for another attempt, of a purely topological explanation this time, arises from these considerations, we leave it here for the moment to turn to some interpretational issues of gauge theories, which will endorse, we believe, our position that a holistic, purely topological explanation of the A-B effect may be the best we can get. The reason for this digression is that one may wonder whether

adopting a different interpretation would provide an adequate explanation within the three approaches we have already discussed.

4.3.3 Active and Passive Interpretations of Gauge Symmetries

As we saw above when we considered gauge transformations, a gauge transformation may be active or passive according to whether we transform-transport the entire physical system changing its spacetime region or we just transform the fields inside the bundle. There we mentioned that mathematically the two are equivalent, yet we said that we need to discuss whether this mathematical equivalence makes any physical sense. In keeping with these two approaches, the gauge fields themselves may be interpreted in either an active or a passive way. For the sake of completeness of the account, let us have a look at the two interpretations and their advantages and disadvantages.

The Active

Interpreting actively the symmetry of a system means that it is in fact the physical system that changes, not the coordinates, and thus one can tell between the different states of the system[186]. In other words, one must actually 'do' something to the system in order to take it from one state to the other. One example of symmetries that receive only active interpretations is that of the discrete symmetries. Take for example the case of reflections. The way to understand this intuitively is by considering that one cannot make her left hand

[186] Of course a symmetry transformation is one that leaves the system unaffected in the sense that one cannot tell the difference between the original and the transformed. However, here we are trying to stress that in an active transformation the physical system *does* undergo some actual change.

coincide with her right hand *unless* one reflects it in a mirror. In other words, the left hand remains left and you can always tell it is left unless you look at it through a mirror.

With regard to the gauge symmetries, on the other hand, when the physical system is in a gauge A_μ does 'look' like a similar system in a gauge $A_\mu + \partial_\mu \lambda$, however, we may accept that the two represent different physical systems or the same physical system in two different and distinct states. Redhead, in his review of Auyang's *How is Quantum Field Theory Possible?* suggested we should adopt an active interpretation of gauge symmetries even when we have to take the holonomies, rather than the gauge fields, as the real physical objects. In either case, we must transcend the observable, which is the electromagnetic field, and consider *the gauge potentials and/or the holonomies as part of the world's basic systems that supervene only on the geometric properties of the spacetime points*. The only price we would have to pay if we considered that either the gauge fields themselves or the holonomies represent some sort of real object on their own would be that then we would have to take on board the existence of some sort of 'metaphysical sub-stratum' in the world, which controls the behavior of the physical, claims Redhead. This increase of metaphysics would not be that bad if it restored locality. But does it? To this question we will return shortly.

The Passive

In a passive interpretation we understand the gauge fields to be some sort of coordinates, so that any transformation that affects them without changing the physical characteristics of the system is just a change of the description, not of the system. Such a trans-

formation, therefore, maps the same physical state of the system to different but equivalent mathematical representations of it. So, coming back to the discussion in the second chapter, we can say that we have ambiguity of representation of the third type, where while the physical system remains the same, there are within the same mathematical structure more than one equivalent mathematical representations of it.

Their problems

The main problem of the attempted active interpretations of gauge theories is the fact that the gauge fields do not seem to correspond directly to something physical, not even when we consider holonomies, hence by considering them as real, one has to cope with an increase in the metaphysics involved in explanations and understanding. Then, a problem that follows is how one could justify the fact that very many of these (meta)physical degrees of freedom need to be eliminated in order to get correspondence between them and physical objects, on one hand, and in order to quantize what needs to be quantized, on the other. To be more specific, in the case of quantum electrodynamics, in order to map the photon to the gauge field A_μ one has to eliminate two degrees of freedom in order to take into account its transverse nature. And even then, physicists have to choose a gauge in order to eliminate the infinite degrees of freedom that are involved and then employ complicated techniques, like the Gupta-Bleuler formalism, in order to quantize it. After all this fuss one is able to calculate measurable quantities, to actually 'measure' the photons.

One might think that the passive interpretation of gauge theories is less problematic than the active one, in the sense that here one does not have to put up with metaphysics.

However, even though we do not have to put up with metaphysics, this is far from being true, because in this case, one has to deal with gauge fixing for two reasons, and this is problematic in its own right. First of all, we want to deal with physical objects or degrees of freedom and it is only a complete set of independent gauge fixed functions that provides one with a complete set of gauge invariant observables[187]. Doughty in his book *Lagrangian Interaction* writes the following about gauge fixing.

> "[T]he existence of a gauge invariance in a system of dynamical equations always implies that one or more of the equations of motion is not a true dynamical equation but a constraint on the initial data. Conversely, equations of motion that contain certain types of constraints on the initial data contain gauge invariances. The choice of an explicit condition to eliminate the gauge freedom of systems is referred to as gauge fixing and the condition is referred to as a gauge condition, which should not be confused with a constraint, although the two are closely related"[188].

And further down:

> "To reduce the second-order electromagnetic potentials to a set which are physical, we must impose a restriction in order to remove the gauge freedom. The new sets of variables will be referred to as being in a particular *gauge* and the restriction is called *gauge-fixing condition*. However, we cannot use an arbitrary restriction which just happens to give the correct number of physical degrees of freedom. Instead we must use only gauge-fixing conditions which lead to new dynamical variables which can be related to the original gauge fields by a gauge transformation."[189].

So, we see that a first restriction in the choice of gauge is imposed by the symmetry itself, as we should expect. But even if we pick up a gauge in accordance with this restriction, and even if in the case of $U(1)$ electromagnetism we are able to do so everywhere, in the case of non-Abelian symmetries we are bound to face the so called *Gribov obstruction or ambiguity*, which does not allow us to choose a single gauge all over the manifold[190].

---

187 For a detailed discussion, see Henneaux & Teitelboim, *Quantization of Gauge Systems*, Appendix 2.A.

188 Doughty, *Lagrangian Interaction*, p.306

189 Ibid., p.398.

190 The so called Gribov Obstruction or Ambiguity was introduced by Gribov (1977) & (1978) and extended

This difficulty arises due to the substantially non-linear character of non-Abelian gauge theories, when one considers appropriate conditions at ∞. What Gribov (1977) showed was that the so-called Coulomb gauge intersected the gauge orbit twice: once at the chosen gauge, as it was expected, and once at a large distance from it. This means that after the gauge has been chosen, the same gauge potential is mapped onto two different, instead of one, gauge equivalent fields A_μ. Shortly afterwards, Singer (1978) put the whole discussion into a fibre bundle perspective and asked whether a true gauge existed in general. By extending the discussion to gauges other than the Coulomb he showed that "topological considerations imply that no gauge exists"[191] when conditions at infinity are imposed.

The second reason is that one wants to be able to quantize the system and a straightforward way of trying to quantize a classical theory like electromagnetism is by quantizing the gauge invariant quantities. It is difficult to do this unless one fixes the gauge because in order "to carry out this quantization, one must find a *complete set of Gauge Invariant Functions...*" [192]. "In practice, it is extremely difficult to find a complete set of observables. Indeed this amounts to solving the differential equations

$$[F, G_a] \approx 0$$

which may not be tractable"[193]. Less difficult, indeed, is to achieve quantization by a different method, that which fixes the gauge by hand! This method works when Gribov obstruc-

---

by Jackiw et al. (1978) and Singer (1978). For detailed discussion of the consequences of it in quantum field theories see Henneaux & Teitelboim, *Quantization of Gauge Systems*, Jakiw et al., *Current Algebra and Anomalies* and Weinberg *The Quantum Theory of Fields*, vol.2.

191 I. M. Singer, *Some Remarks on the Gribov Ambiguity*, Commun. Math. Phys., vol.60, 7-12, (1978).

192 Teitelboim & Hennaux, *Quantization of Gauge Systems*, p.275

193 Ibid.

tions do not prevent us from fixing the gauge globally, and it simply consists of imposing canonical gauge conditions

$$\chi_a = 0.$$

This is legitimate because any function of the canonical variables can be viewed, after complete gauge fixing, as the restriction in that gauge of a gauge invariant function. Hence, once the gauge is fixed, one is effectively working with gauge invariant functions. Furthermore, one finds that the Dirac bracket associated with the constraints ($G_a = 0$) and the gauge conditions ($\chi_a = 0$) is just the Poisson bracket of the corresponding gauge invariant functions, so that the Dirac bracket yields the correct bracket in the reduced phase space. "With canonical gauge conditions, the reduced phase space quantization becomes identical to the quantization of the 2nd class constraints" [194], because after the conditions have been imposed, the symmetry is gone and the constraints that remain -including the gauge fixing conditions- behave as second class.

However, *the gauge fixing* or *reduced phase space approach* may suffer from drawbacks other than the Gribov obstruction. The elimination of the gauge degrees of freedom -i.e. the fixing of a complete set of gauge invariant observables- may spoil manifest invariance[195] under an important symmetry and hence one may lose important information. Moreover, the brackets of the complete set of observables that one has found may be complicated functions of these observables, and their quantum mechanical generalizations may not be straightforward. Similarly, the Hamiltonian in terms of the independent degrees of

194 Ibid., p.276.

195 *Manifest* here means *linear*.

freedom may turn out such that it is impossible to give a quantum mechanical definition of it.

Of course, there are other ways to proceed to quantization like for example the Dirac approach where the gauge degrees of freedom are not eliminated, or the Dirac-Fock approach where the constraints are implemented differently[196] which fix the gauge at the end. But even within these approaches the problems abound. In the first one, for example, the fact that the gauge degrees of freedom are not eliminated entails that the representation space carries information that does not correspond to anything physical and hence further assumptions are required; by doing so, Dirac's approach and the reduced phase space method are formally equivalent, hence the problems that infect the first are present in the second as well. As for the Dirac-Fock approach, the price one has to pay there is that some of the resulting operators produce states that do not correspond to anything physical.

The conclusion that follows from this discussion, then, is that within an active interpretation of the gauge theories, the gauge fields acquire the status of physical objects, but then more metaphysics is involved in the explanations. The problem of indeterminism is not solved, local action may be satisfied, but separability is not. As for the passive interpretation, in it gauge fields have to be eliminated either using gauge fixing - which in the theories we are concerned with cannot be done due to Gribov obstruction- or by some other mathematical manipulations of the theory which involve their own problems. Non-locality cannot be avoided here either and the problem of indeterminism depends on whether the

196 For more details see Henneaux & Teitelboim, *Quantization of Gauge Systems*.

one gauge can be found -in the first and Dirac's approaches- or is overshadowed by the existence of non-physical states -in the Dirac-Fock approach.

After the discussion about active and passive interpretations of gauge theories the question that remains open is whether different interpretations impinge on the attempted explanations of the A-B effect. If we adopt what Lyre calls the A explanation, the problem is non-locality due to violation of local action. In this case, adopting the passive interpretation we have unequivocal violation of local action, as Healey showed. One would expect that adopting the active interpretation one would manage to get around this difficulty, and this is what Redhead anticipated. But if we give it a second thought, we realize that although adopting an active interpretation remedies the problem of supervenience, it does not guarantee local action, because the crucial point is not only whether the tentative physical entities supervene or not on geometric properties of spacetime points but also whether they are where they should be, namely along the path of the electron. If the 'one true gauge' was the one suggested by Healey, then non-locality is still present. But this very idea of the existence of 'one true gauge' is loaded with metaphysics since there is no physical necessity that dictates its existence nor any indication that there might be. It is inspired, rather than dictated, by the wish to solve the problem of determinism of gauge theories, but determinism does not need gauge fixing; let alone that the requirement of determinism itself is more of an assumption than of a physical necessity. Moreover, if such a thing as the 'one true gauge' existed, it should be defined all over the manifold at once because fixing the gauge means picking up one out of the infinitely many divergencies that comprise the gauge trajectories. This, in fibre bundle language means choosing a cross section and this

has to be done all over. But the Gribov problem makes it impossible, as we have already mentioned. Hence, even if one was willing to pay the price of increased metaphysics, one has not established the sought after locality. So far as the C approach is concerned, whether we choose active or passive makes no difference to the problem of non-separability. The very fact that the physically significant entity is a loop implies that one should not expect explanations involving separable processes.

We may now conclude that none of the suggested interpretations and approaches managed to solve the problems raised. However, to our view, the elusiveness of locality constitutes no problem at all since it only points towards a holistic explanation, where the gauge field is not perceived to be a localized causal agent any more. Its role, as we shall see, is that it informs us about the interactions that occur in the physical system. Finally, let us remark that after this discussion, the reason behind our anticipation, in chapter two where we discussed the surplus structure and the ambiguity of the third type, of passive interpretations of gauge symmetries becomes clear: an active interpretation of gauge symmetries not only would solve none of the problems but it would also increase the metaphysics. And now we may proceed to discuss a fourth way to the A-B effect.

4.4 A 4th Way to the A-B Effect

The fourth way to the A-B effect provides a holistic explanation of the phenomenon. This kind of explanation does not fit any of the D-N, C-R or unification models of scientific explanation. As it takes into consideration the entire system rather than small parts of it causally related to each other, one naturally wonders if it fits the model of teleological

explanation. Here we show that it does not. Hence the explanation of the A-B effect stands as a distinctive kind of explanation, which we call *topological.* But let us state the explanation first, and then see how it does not fit any of the aforementioned patterns although at the same time it does have certain characteristics that partly match them.

4.4.1 Holistic Approach in a Topological Explanation

The fourth explanation about what is going on in the A-B effect is based on topological considerations and one may find very good reasons for both liking it and not liking it. It is the approach favored by many mathematical physicists[197] and we were directed towards it for several reasons. The fact that there does not seem to be a satisfactory bit-by-bit causal account of the phenomenon, which is the result of the non-separability present in any other attempts to explain the effect, indicates that we should take more into account than just the (speculated) events and physical processes along the path of the electron. Knowing what is going on in some parts of our physical structure is not enough since this knowledge leaves out pieces of information that cannot be retrieved. Therefore we require a formalism that contains all the necessary information for a good comprehension of the events. This formalism, we suggest, is the fibre bundle formalism, in which the mathematical entities of the surplus structure *register* all the information -not just bits-and-pieces of it- about the topology of the base manifold. Consequently, the mathematical objects involved do not dictate the behavior of physical objects as though they were the *causal agents* acting on those physical objects, nor they are held responsible for a signalling process that takes place

[197] For topological accounts and explanations of the effect see, for example, Nakahara, *Geometry, Topology and Physics*, Nash & Sen, *Topology and Geometry for Phycisists*, Ryder, *Quantum Field Theory.*

-allegedly- between solenoid and electrons. Instead, they are descriptive tools that encode all the information of the properties of spacetime and for this reason they account for the effect in terms of the relations between the spacetime points and the physical objects, i.e. the electrons, involved. From this perspective one could say that the solenoid has modified not just the spacetime points that it occupies, but also the region around it. The shift in the phase of the electrons happens because the spacetime points along its trajectory are thus modified. The gauge field does not participate in this modification, it just encodes it and it gives us a mathematical tool that allows for measuring the results this change brings about. A measure of the results is provided by the holonomies. In this way, we gain full *awareness* of all the elements involved and the factors affecting the electron and a good understanding of its behavior. But let us examine how this is done.

Holonomies, Homotopy and the U(1) Group of Electromagnetism

As we have already seen, the fibre bundles involve mappings between a base and some other manifold and these mappings carry all the information about the structural characteristics, or the topology, of these spaces. The discussion there is related closely to the discussion on the topological non-separability of the A-B effect and the holonomies that are involved, and, among other things, it is quite revealing about the relation between mathematics and physics. We will leave the discussion for the relation between physics and mathematics for the next chapter but let us explore here how the discussion on loop integrals in topologically non-trivial manifolds fits in the more general picture of the fibre bundles and how it relates to the A-B effect.

One account that aims at explaining the A-B effect could be the following. Assume that the base space in our discussion is the spacetime manifold with a solenoid in it. For the sake of simplicity, we can consider a slice of it, which is described mathematically as a plane with a hole. The hole represents the area occupied by the solenoid that is inaccessible to the electron. At the same time, the presence of the hole renders the configuration space topologically non-trivial or, in other words, not simply connected. This only describes the fact that the hole is a region that the electron cannot access. The infinitely many curves surrounding the solenoid are equivalent in the sense that they can be deformed into each other continuously, but they cannot become zero. We say that the functions representing these curves are *homotopic* -i.e. map preserving- and they belong to a group called the *fundamental group* or *first homotopy group*. The functions describing the curves have parameters that take values from the interval $[0, 1]$. Hence, this space, call it X, topologically corresponds to the direct product of the line $\mathbb{R}^1$ and the circle S^1, namely $\mathbb{R}^1 \times S^1$. The electromagnetic field that is involved in the origination of the phenomenon is a physical entity that is described using the $U(1)$ group G, and the topology associated with our group is also that of the circle S^1. A fibre bundle is generated by the base manifold and the group and its structure is as was described above. The connection in this fibre bundle is the field A_μ[198] and the electromagnetic field is represented by the four-dimensional curl of A_μ which is also known as the curvature. Given that the actual magnetic field, or curvature, is zero everywhere on the manifold, we are talking about vacuum here, where the curvature is zero, but the connection not necessarily so.

[198] The connection follows the general transformation rule $A^\mu \rightarrow A^\mu + \partial^\mu \chi$. Because in our case we are in vacuum, we can write that $A^\mu = \partial^\mu \chi$.

Ryder[199] writes that "the gauge function χ is a mapping from the group space G onto the configuration space X: $\chi : G \rightarrow X$ whose non-trivial part is given by $\chi : S^1 \rightarrow S^1$". In the terminology we have introduced above, this means that this is a connection one-form pulled back to our base space. We have already said that the fibre bundles as formalism are so structured that all the information about the topology of the base space is included in the structure of the bundle space and vice versa. Here we can see how this is realized in the A-B case, where the non-trivial topology of the base space is reflected by the non-trivial topology of the group used to define the principal bundle. Ryder argues that the fact that the electromagnetic field is zero outside the solenoid, along with the fact that the gauge field χ is not, entail that χ is not single-valued. If χ is not single-valued then the G space is non-simply connected. If χ was single-valued then the loop integral would be zero. But the loop integral is not zero, hence χ is non-single-valued and therefore G is non-simply-connected. And hence, he concludes, "it is an essential condition for the A-B effect to occur that the configuration space of the vacuum is not simply connected"[200], where the term vacuum refers to the absence of magnetic field in the configuration space outside the solenoid. Along the same lines was Lyre's conclusion, as we have already seen. But in both cases, the necessity they argue for does not follow. Only as an assumption or a crude induction one could claim that the electromagnetic field is zero *and* at the same time the gauge field is not zero *only if* the topology of the base space X is non-trivial; for the topology of the base space in the case of the A-B effect is trivial indeed: the presence of the solenoid does not

199 Ryder, *Quantum Field Theory*, p.107.

200 Ibid., p.105.

create a hole in spacetime. One might claim that, nevertheless, the following approximation provides a valid topological explanation of the effect.

Topological Explanation (1)

There are two variations of what we may consider as a topological explanation of the A-B effect. First of all, one notices that the difference in magnitude between the electron and the solenoid is of the order of 10^{10}. Given that the energies we are talking about are very low, this means that a very big chunk of space, $10,000,000,000$ bigger than the electron itself, cannot be accessed by it. So, from the perspective of the electron, *it is as if* spacetime is topologically non-trivial where the solenoid is, and that might be considered as a very good approximation. Moreover, even from our point of view, treating the space outside the solenoid as topologically non-trivial is not a far-fetched idea if one considers the limiting case where the solenoid is shrunk to a point. The point-solenoid cannot be made to disappear completely and hence one has to accept that the spacetime manifold is not simply connected. Non-simply connected manifolds have non-vanishing holonomies, which means that the parallel transport of a matter field along a closed curve that surrounds the 'hole' results in a shift on the phase of the field. One then could claim that the reason for the shift is that spacetime has been modified as a result of the non-trivial topology and the description of -or the information about- this modification is given -or encoded- by the gauge potential; the potential, though, does not cause the shift. Therefore, an explanation of the phenomenon involving non-trivial topology that entails non-vanishing holonomies might be appropriate since anything else -that is, the zero magnetic field or the non-physical

gauge potentials - would not adequately describe what is happening there. From the prespective of the fibre bundles, the non-trivial topology of the base space is associated with a non-trivial bundle space where a cross-section cannot be defined continuously all over it. So the connection on the principal bundle changes as we move around the solenoid and the consequence of it is that the phase of the matter field -defined on the associated bundle- changes as well.

One important point to clarify here is that what is really important for the effect to happen is not just the material presence of the solenoid in the set-up, for one then might claim that even when the solenoid is switched off the region inside it is still inaccessible to the electron and yet there is no A-B effect. What is crucial for the effect to happen is the flux of electromagnetic field inside that apparently modifies the connection of the spacetime around it and one could assume that this modification takes place in a way that is in accord with relativity theory. Hence we might approximate the inaccessibility due to the presence of a solenoid with a magnetic field in it with a spacetime which is topologically non-trivial.

Topological Explanation (2)

The topological explanation of the effect may be given a different gloss. One may assert that in this case it is not the presence of the solenoid that makes the topology of $\mathcal{M}$ non-trivial, rather, it is the topology of the bundle vacuum itself -and hence of the configuration vacuum- that is non-trivial and as a consequence the phase of the electron-field is shifted as it passes through, where vacuum in this context is defined as a region

where the energy of the electromagnetic field is zero. The connection of the principal bundle -that is to say, the gauge field A_μ- describes how the phase shift occurs and it is not the causal agent responsible for the shift but an information bearer instead: it just contains all the information about how the matter fields should behave as they move along the spacetime manifold. The curvature of the total space is nothing other than the familiar electromagnetic field, which cannot be considered to be a causal agent either, as we have seen. Instead, it may be regarded as a property of the spacetime points, conferred to them by the modified topology of the base manifold.

This version of the fourth way differs from the second, or *A*, explanation of the A-B effect because here we do not need to rely on the reality or the locality of the gauge field. What matters in this case is the non-triviality of the base manifold which affects the bundle space by changing the value of the connection in it and this describes a change, a shift, to the phase of the matter field. Once again, one is able to tell a story about how this modification occurred that is perfectly compatible with relativity principles. Moreover, since we do not need to rely on the reality of the holonomy either, it differs from the *C* approach as well: it is the topology, rather than the holonomy, which constrains and controls the effects on the physical objects. As the explanation we are considering here is purely topological, we do not need to consider the holonomies as the fundamental causal entities either; it suffices to say that the non-trivial topology of the vacuum, which results in phase shift or non-vanishing holonomies, accounts for the effect and, once again, the holonomies are merely a measure of the effect. Hence in this way of explaining things we obtain a holistic causal picture where the ultimate 'cause' of the shift is the topology. The

modified topology endows the spacetime with some properties, which in turn affect the physical objects that move around in it. The importance of the fibre bundle formalism is that it provides a complete tool for the precise description of the phenomenon and for the calculation of quantities that are measurable.

We said at the beginning of this section that there are several reasons why one may or may not like the approach we just presented. First of all, and before we actually assess the topological explanation, we would like to mention two possible objections to -or reasons for not liking- it that would persist even if the topological explanation turned out to be a *bona fide* explanation. The first one is that we give up completely the idea of ever getting a local causal account -at least within this formalism- while the second is that we also part with determinism in the sense that since up there, in the bundle space, we have more entities than down here, there are infinitely many gauge fields corresponding to one electromagnetic field, hence starting from well defined initial conditions, we may end up in one out of infinitely many possible final states of the total space. But if this is a problem, then it seems that it is inherent to the way physical objects are represented by mathematical entities, at least within the context of gauge theories. Remember the discussion in the third chapter about what we called ambiguity of the third kind, which seems always to be present in this type of physical theory.

4.4.2 Teleological and Topological Explanation

Is this holistic explanation a teleological explanation as well? If we regard as teleological the type of explanations that we discussed previously in this chapter, the topological

explanation would also be teleological provided that the system under consideration was a directively organized system, that is if it satisfied the four requirements set by Nagel. The first three assumptions are more or less satisfied if we consider the following correspondences. If we take the spacetime manifold and the electrons that move in there as the causally relevant parts of the system, then the first assumption is satisfied. These are independent in the sense that we could change either of the two without an immediate necessary change in the other; for example we could change the properties of the manifold or the number of the electrons independently from each other. However, if we vary the topology of the physical structure, then this would result in an adaptive variation to the behavior of the electrons; hence the third assumption is also satisfied. So the issue in this case is whether the fourth assumption is also satisfied. As a matter of fact, it is not, and here is the reason. According to the last assumption, *the values that the primary variation has assigned to the initially changed variables correspond to the values the adaptive variation has assigned to the adaptively changed variables so that S is eventually in a G state again.* But this assumption is not satisfied by the A-B set-up and its states. The initial state of the system is a state with the electrons on one side of the solenoid with a certain phase, while the final state contains the same electrons in some other spacetime location with a different phase. So even if the spatiotemporal coordinates of the physical entities were not considered as independent variables, their phases should. Hence the system undergoes an adaptive variation that does not take it back to its initial state *G* and, therefore, our holistic explanation does not fit Nagel's idea of teleological explanation.

Nonetheless, although Nagel's fourth condition seems to be essential in biological systems that are sustainable only when a change in their state is followed by adaptable processes that will return the system in its previous state, it does not seem necessary in a physical system like that in an A-B setting. The behavior of the electrons in such a system may be considered to be goal-oriented, where the goal is the electron's phase shift while the reason, the cause we dare saying, behind the shift, is just the topology of the base space or the vacuum. This way one may explain why -but not how- the shift occurs. When we discussed Nagel's teleological explanations we mentioned that in his account he tried to avoid any reference to final causes, because physicists do not like their explanations to rely on such obscure metaphysical notions. With our suggested modification of Nagel's account have we managed to avoid such references? Given that the topology of the bundle space for the *U(1)* group is non-trivial, if the topology of the base space turned out to be non-trivial as well, we would have good reasons to claim that our suggestion constitutes an explanation free of metaphysical considerations. But if the base-space manifold is trivial, as we will argue in a while, our acceptance of a teleological explanation would rely heavily on metaphysical assumptions. Hence a claim about the goal-orientation of a system like ours is one loaded with metaphysics and we do not want to commit ourselves to it, especially since it does not serve any purpose.

4.4.3 D-N Model and Topological Explanation

According to the D-N model, an event is explained by subsuming it under general laws. The explanation is a valid argument, the premises of which are those general laws and

statements describing particular facts. In our case study, the explanation we offer is definitely not of this type. The claim is that what is responsible for the effect is held to be a certain change in the topology of the spacetime manifold, and this is clearly not a law-like statement. On the other hand, one could not claim lightheartedly that it is a fact either. As we shall see shortly, we may consider it to be, at most, an idealization concerning the boundary conditions. In the full explanation of the effect we definitely rely on law-like generalizations. One is that all interactive physical systems are described by Lagrangians that are invariant under variations at the boundaries. Another one is that all the fundamental interactions in nature arise when we require that the actions describing the physical systems are gauge invariant. The fibre bundles formulation of gauge field theories is a perfect deductive system. But although we take these two generalizations and the equations of motion of the fields to be true, they do not explain the effect by themselves. The topological considerations, on the other hand, though they may be formulated as a general statement, they are specific to each particular problem and hence do not qualify as laws. Moreover, the theory as a whole involves gauge fields -our connections- that play an eminent role in the derivations, yet they do not take specific values. One could claim that since the treatment so far has been classical and since it is only gauge invariant quantities that really matter, the gauge fields are only used in sub-derivations so they do not spoil the deductive character of the explanation. Nevertheless, one should bear in mind that the main purpose of these theories is the study of relativistic quantum fields and it is explanations involving these kind of fields that we try to assess here. In these conditions, then, the connections do participate in the explanations not as auxiliary assumptions, nor as causal agents, but definitely as part

of the ontology and since they cannot be attributed a definite value, certain statements that include them -like for example the gauge fixing conditions- cannot be given a definite truth value.

4.4.4 C-R Model and Topological Explanation

The C-R model advocates that by citing the causally relevant factors and mechanisms that are responsible for the phenomenon we explain it. The three previous attempts to provide an explanation for the phenomenon were doing precisely that, they were seeking for legitimate causal mechanisms. The underlying assumption in all these attempts was that the causally relevant factors act locally. But as we saw, all these attempts failed. In our fourth explanation, one of the main aims was to avoid precisely the use of any dubious causal mechanisms in it. Hence this explanation, though it may involve causal relations and mechanisms, it is not a C-R explanation.

4.4.5 Unification and Topological Explanation

The theory that supports the topological explanation of the A-B effect is that of electromagnetic interactions and, as we saw above, it is part of a larger family of physical theories, namely the theories of the fundamental interactions which are mathematically formulated using the structure of the fibre bundles. The fibre bundles provide all the mathematical tools we need for the description of fundamental interactions -along with some surplus structure, which in the case of electromagnetic interactions we had some difficulty in interpreting as physical. However, from the perspective of our topological explanation it is this

very surplus structure that provides a full description of the new properties of the spacetime manifold, which are due to the presence of a solenoid in it or, mathematically speaking, due to its non-trivial topology; and it is this description that tells us not just that the shift occurs, but also what its magnitude is. In this explanation, one cannot consider the A-B effect to be a mere consequence of the bigger unified picture because the fibre bundle formalism only tells you that all the information about the topology of the base manifold is contained in the bundles as well in a specific way, that is using the principal bundle. It also tells you that all the information about the matter fields and their whereabouts is contained in the tangent bundle. But there is nothing said about the particular situation we face when we examine the phenomenon. Hence, the bigger, unifying picture puts the phenomenon into a larger perspective, but it does not explain it; at least not on its own. On the other hand, as we observe things in this bigger picture we realize that this unified approach is revealing about the relation between the mathematical and the physical: the connections control formally -but not causally- the physical in the sense that accurately describe what is happening there.

4.5 A First Assessment of the Topological Explanation

One thing that arises from this discussion is that in topological explanation we use elements from all the models of explanation we have discussed, namely teleological, D-N, C-R and unification. Yet, the explanation stands in a category on its own, thus we could maintain the special name *topological* explanation. One might argue that we could give it the name non-local or holistic instead. A closer look at it, though, shows that this explanation is not

really non-local in the sense that the actual topology is described locally and there is no kind of action at a distance involved in it because the entities of the theory that could be held responsible for non-locality either do not play a causal role or they are not needed at all.

The topological explanation relies on laws and derivations from them, contains references to causal elements, and the particular events that we examine may fit in a more general unified theory; but there are also two more things in it than just these. First of all, we have to take into account the entire physical system, not just what we might consider to be the assembly of 'causally relevant' elements of it -hence it is holistic. The reason why we prefer the name topological rather than holistic is that although it is holistic there is more to it, namely the consideration that the actual effect takes place because of a change in the topology. Second, we use a mathematical structure, which although it seems to represent the physical entities involved along with a whole lot of surplus structure, as a matter of fact it minimally encodes all the information of the entire system, albeit using some entities topological in nature that may not correspond directly to physical entities; nevertheless, these entities, as the objects that encode the entirety-of-information, dictate the behavior of the physical. Do they govern it? No! But we do not see a problem in it because in physics we do not necessarily use the ultimate causes in order to explain physical events. Often, we only look for information that may reveal possible causal links between the objects involved and theories that help us predict behaviors as well as measurable quantities. In our case, gauge theories and their formulation in terms of fibre bundles do both, very successfully indeed. Encoded in the form of the gauge fields -or connections- is all the information

about how the base space has been modified due to the presence of sources and hence those fields reveal the link between their presence and the change in the behavior of the electrons, while at the same time the predictive power of the complete theory has been proved to be overwhelming.

This theory with its double success links the physical (i.e. everything that happens in the actual world) with the mathematical (i.e. a lot of information -if not all- about the physical objects and their relations is contained in here) and uses experiments and measurements to validate this relation. To our view, one should seek the very deep connection between physics and mathematics in here, in the fact that once a theory is formulated in a mathematical language, it provides measurable properties and it allows for quantitative inferences and measurements. But some further elaboration of this point needs to wait until the following chapter. In the mean time, one is more than justified to ask: does the claim that the topology is non-trivial provide a *deep* explanation? If by 'deep' we mean an explanation where all the factors involved are known and all the statements are true, then the answer is *no*, at least so far as the A-B effect is concerned; for, to begin with, the topological claims in the attempted explanations of the A-B effect are not true.

4.5.1 Assessment of Topological Explanation (1)

So far as our first attempt is concerned, there is a crucial disparity between the alleged approximate explanation of the A-B effect and the legitimate approximate explanations that were discussed previously in this chapter. In this case, like in the case of chaos theory, we make use of a model that clearly involves a negative analogy between the model we use and

the physical system we aim to describe, namely we consider that a spacetime manifold with a solenoid in it is non-trivial. However, unlike the chaotic examples, here we require from this very analogy to causally explain the physical events, hence it is essential since its non-inclusion would undermine even the positive analogies. From Hesse's perspective, the only reason we would have to accept this explanation is that we have no better alternative. There is a striking success of this type of explanation, though, that makes one wonder whether there is a slightly different, legitimate, way of accounting for the effect. The success is that when used as a formal analogy, it predicted the weak vector currents and gave rise to the unified theory of the electroweak interactions, and one guess for the different account might be what we called topological explanation (2).

To conclude this section, we would like to state clearly that tempting though the approximation may be, it does not constitute a legitimate explanation. Yet, at the same time, there two things in this account that we should bear in mind. The first is the fact that the holonomies are non-vanishing. Although this is not a necessary condition for explanations that involve nontrivial topologies, it is a good indication that there is something about the electromagnetic field that points towards explanations that are holistic in character. On the other hand, the topological considerations that are sufficient for non-vanishing holonomies provide very far reaching heuristic, or formal, analogies.

4.5.2 Assessment of Topological Explanation (2)

The vacuum state that this interpretation of the topological explanation requires is a state where the electromagnetic field is zero. The fact that there is a solenoid with electromag-

netic flux inside in some finite region of spacetime means that one could consider that vacuum extends over the rest of spacetime except from the region occupied by the solenoid itself. But surely, in this second attempt to provide a topological explanation, the alleged vacuum is not really a vacuum due to the presence of the solenoid and therefore things seem to be at least as bad as in the previous attempt because although now one might consider the claim that there is a vacuum outside the solenoid as true, the fact is that vacuum in quantum field theories is a global state of the field. This fact does not allow for any concessions because if the state was really a vacuum state, then the global vacuum would imply local vacua. However, the presence of electromagnetic field at some region of spacetime spoils the vacuum state altogether and no notion of approximation can save it. It seems, therefore, that once again our attempts to salvage the topological explanation of the A-B effect using approximation have failed.

The situation we encounter in the explanation of the A-B effect could be compared to the classical case of projectile motion[201]. In projectile motion, in order to explain the parabolic trajectories, one has to ignore the 'accidental' frictional forces and to assume that the gravitational field strength g is constant throughout the path of the projectile and with direction perpendicular to the surface of the flat earth. So, one considers the curvature of the earth to be zero, locally, and hence one changes its global topology from that of a sphere to that of a plane. In both the A-B and the projectile cases, we have exchanged the actual topology of the physical system with a different one and we therefore use a negative analogy for explanatory purposes. At the same time, in the A-B case, as well as in the

[201] This analogy was an idea of Professor M. Redhead, to whom I am grateful for it.

gravitational, it is not the change in the topology that provides the deep -that is to say the true causal- explanation for the phenomena, rather it the presence of the solenoid in the former and that of the gravitational field in the latter.

Nevertheless, one may claim that there is a major difference between the two approaches: in the A-B case either there is or there is not vacuum, while in the projectile motion case the change of topology may be thought of as just an approximation where the gravitational field lines are approximately parallel lines and the surface of the earth is approximately a plane, therefore the trajectory is approximately part of a parabola. The argument goes then that in the case of projectile motion we just approximate the actual physical situation with some mathematical structure that does not essentially misrepresent it and this is because the negative analogy in this case does not causally affect essential properties of the system. The truth of the matter, though, is that the negative analogy does affect the essential property that the gravitational field strength is inversely proportional to r^2; and the conclusion that follows is that although we might consider a gravitational field with parallel lines near the surface of the earth as a good approximation, the alleged change in topology fails to serve any explanatory purposes. In both cases, then, by using topological considerations one exceeds by far what one might consider as reasonable limits of approximation and idealization. Yet, in both cases we get useful and fruitful -in an explanatory sense- insights about the relations between the physical objects involved in the processes, while from the formalism as a whole we get very good predictions about their future behavior and certain measurable quantities.

Are we justified to say that a topological explanation like the one we employed for the A-B effect misrepresents reality? Literally speaking, yes we are. For one reason, the base space manifold is trivial despite the presence of the solenoid in it and for another, the vacuum is not really a vacuum for exactly the same reason. However, this 'failure' of the non-trivial topology of the mathematical structure to 'explain' the physical events is not a sufficient reason to reject the theory or to undermine its *heuristic* power. In the following chapter we will discuss again and at some length the notions of idealization, approximation and abstraction that are involved in scientific explanations in general and in topological explanations in particular and we will see then that although not true, and hence not a good explanation from this perspective, the topological account of the A-B effect is a very useful device for other reasons.

Things, however, take a different turning in relativistic quantum field theories because, as Redhead showed (1995a), (1995b), the straightforward relation between the global and the local vacuum state that we mentioned above breaks down in there. Of course once again we make a leap and starting from a classical discussion we draw conclusions about relativistic quantum objects, but we are justified in doing so because whatever we have discussed so far applies in the quantum case as well and because we are not really interested in what is going on in the classical cases only; these just provide a stepping stone. What could we say then about the topological explanation (2) of the A-B effect in the case of a relativistic vacuum, where a global vacuum state does not prevent observables from exhibiting quantum fluctuations? Since "these vacuum fluctuations of local observables are a charac-

teristic feature of the *relativistic* vacuum"[202] one is justified to claim that in the A-B case the state of the field is indeed a vacuum state despite the fact that locally it takes non-zero values. To take the old Aristotelian line of argument, one could claim here that the vacuum state of the relativistic quantum fields is not space(time) empty of objects. Rather, it is a field defined over spacetime that allows for either manifestation or not of observables, locally, due to its quantum fluctuations. Hence a vacuum state that is compatible with the presence of objects in it is reminiscent of Aristotle's wooden cube immersed in the water, only in this case the water-field penetrates the cube-solenoid throughout its extent and so interpenetration and therefore coexistence become possible[203].

We feel compelled at this point to stress that the main aim of gauge theories is to describe elementary particles and the fundamental interactions, both of which are quantum and relativistic physical entities, in a unified way, if possible, and to a great extent they have done so. In these attempts, topological considerations and non-trivial topologies are used as positive or neutral analogies and play a fundamental role in explaining as well as in probing the theories.

4.5.3 Topological Solutions

The above discussion about the vacuum state of fields and the possibility of a base space with a non-trivial topology become legitimate and worthwhile reflections when one considers stable extended solutions to the Euler-Lagrange equations of motion of non-linear field

---

202 Redhead, (1995b).

203 For detailed discussions about vacuum see Aristotle, *Physics*, Jammer, *Concepts of Space* and Grant, *Much Ado About Nothing.*

theories. The Yang-Mills theories are non-linear and the topological solutions offered are well defined topological objects with finite energy, which have the general name *solitons*; *monopoles* and *instantons* -or *pseudo-particles*- are soliton solutions too. Soliton solutions have been given serious thought by theoretical physicists over the past twenty five years or so because they sidestep the problems of infinities and renormalization; these problems impair quantum field theories that describe basic matter fields of nature as though they were point objects. However successful these theories of point-objects may be, the quest for something more satisfactory continues and the stability and finitarity of the topological solutions has been very promising, in terms of the explanations it provides, and alluring, so far as its heuristic powers are concerned.

The first one to introduce the term monopole was Dirac (1931) and his main incentive was to remedy Maxwell's equations from an apparent asymmetry: though they allow for electric charge, they do not allow for magnetic charge in the form of magnetic monopoles. By introducing a radial magnetic field, Dirac made the equations symmetric and arrived at the monopole solutions and the quantization condition of the electric charge that is guaranteed by the presence of magnetic monopoles. In the case of electromagnetism, where the symmetry group is *U(1)*, although the presence of monopoles makes it more symmetric between electricity and magnetism, their very presence is not necessary. Hence, the existence of magnetic monopoles is not determined -not even on this theoretical level- by the possibility that they can be accounted for by the theory. However, in the case of Yang-Mills gauge theories, especially when spontaneous symmetry breaking is introduced, there emerge solutions to the field equations -the Higgs fields- with magnetic charge, despite the

fact that the only charges present in the matter fields of the theory are electric. So, where does this magnetic charge come from? The origin of such magnetic charge, or rather of such magnetic monopoles, is topological and their theoretical possibility was discovered by Polyakov (1974) and 't Hooft (1974). The main idea behind them is this. Both the Yang-Mills action and the Euler-Lagrange equations are non-linear and for a theory with gauge group *U(n)* they take the general form

$$\begin{aligned} S &= \frac{-1}{2}\int_M tr F^{\mu\nu}F_{\mu\nu}dv \\ [D_\mu, F^{\mu\nu}] &= 0 \quad (a) \end{aligned}$$

or in terms of two-forms

$$\begin{aligned} \mathbf{S} &= -\int_M tr\mathbf{F}\wedge^* \mathbf{F}dv \\ \mathbf{D^*F} &= 0 \quad (b) \end{aligned}$$

respectively. The Euler-Lagrange equations (a) and (b) are non-linear equations containing quadratic and cubic terms in **A**, the connection, and in general they are not solvable. However, if there is a connection such that

$$\mathbf{F} = \lambda^*\mathbf{F}$$

for some With these conditions, the map

$$g : S^3 \rightarrow SU(2)$$

falls into homotopy classes or, in other words, every g is labeled by an integer k, which is called the degree of g and classifies principal bundles with group *SU(2)* over S^4. S^4due to the boundary conditions may be considered as a non-contractible sphere made of two

overlapping and contractible hemispheres. These mappings, or hemispheres in our case, are not continuously deformable into one another and hence they are topologically distinct. In the areas of overlap A_μs are related through gauge transformations. Hence, the integer k labels both asymptotic data of $A_\mu(x)$ and the bundle P to which $A_\mu(x)$ belongs. The result is that the topology is no longer trivial and the soliton solutions that emerge carry magnetic charge. Abelian, as well as non-Abelian monopoles are constructed in a similar manner. One very important non-Abelian monopole is the Yang-Mills-Higgs monopole whose discovery or not will determine whether the so-called standard model is really viable.

With their reformulation of Dirac's theory using fibre bundles, Wu and Yang (1975) revealed the similarities between Dirac's idea and the monopoles in the non-Abelian gauge theories, as well as their differences. The main difference between them is that in the *U(1)* case monopoles are inserted into the theory while in the non-Abelian cases they become a necessity once the boundary conditions are set. A very important feature of these solutions is that they are stable and their stability is a result of the fact that the boundary conditions fall into distinct classes, those labeled by k, only one of which corresponds to the vacuum state that is, of course, global and degenerate. The fact that they are stable and with finite energy makes these mathematical objects very appealing because they do not run into the infinite-energy problems that the point-entities we nowadays identify with the elementary particles do, hence renormalizability is rendered irrelevant, and therefore they may be proved to be the 'real' fundamental entities of nature. Moreover, quantization of the electric charge would follow from that and the quark confinement would be accounted for. So, if nature concedes to this view by giving us some experimental evidence that monopoles

exist, the far reaching topological explanations will prove to be indispensable, very good explanations with true premises and, therefore, true conclusions.

4.5.4 What More There Is in the Fibre Bundle Approach?

The attempt to explain the A-B effect is just a simple example which illustrates what one may do with a formalism as rich as the fibre bundles. However, there are a lot more possibilities in this formalism and we would like to give a brief account of some of them in this section.

One very basic assumption in physics is that we observe fundamental fields through their interactions, therefore any theory that purports to describe these fields must allow for their description. Gauge theories describe interactions successfully and when examined form the fibre bundles' point of view, they give a unified picture of all the known interactions. The thing with the fibre bundles is that they allow for many possibilities, infinitely many as a matter of fact. With the idea of the fibres over each point of the base -or spacetime- manifold, it is as if a whole new world opens up over every single point: a world that describes what is happening on the base manifold by using the plethora of the tools available in it but not in the base manifold. Moreover, all the information can be readdressed and conveyed back and forth.

The coupling terms, which can be used to describe interactions, arise when we require certain theories to be invariant under specific symmetry transformations. In this case, just by using variational techniques we get equations for both the matter fields as well as the fields with which they interact. The matter fields in the fibre bundle formalism are

represented by tensor fields -which are cross-sections on the tangent bundle- while the interaction-carriers are viewed as connections on the principal bundle, with which the tangent bundle of the matter fields is associated. Thus we express interactions in a unified and coordinate-free way while at the same time we get a clear distinction between the matter and the interaction fields, which we would expect to be different. This theory can accommodate electromagnetic, weak and strong interactions as well as gravitational interactions -though the latter are somewhat different and in the case of the weak and the strong interactions further properties of the interactions require some modifications of the theory[204].

[204] Here I am referring to short-range of the weak interactions -which led to spontaneous symmetry breaking- and to the quark confinement. However, I will not discuss these issues here, because they fall beyond the present purposes.

Chapter 5
Conclusions

In this final chapter we will try to pull together everything we have discussed so far including all the historical information we have presented and some further philosophical insights. The goal of this thesis is an extended exploration of the relation between mathematics and physics and we attempted to address the issue from two perspectives, one historical and another philosophical. Our main conclusion from the history of gauge theories and fibre bundles was that although the mathematical theory developed quite independently from the physical, there was a strong physical intuition that was at its very heart. Was that physical intuition, then, what made the mathematical structure so relevant to the world? Yes but not on its own, for there is also the process of abstraction involved, the inevitable route that takes us from the world as we experience it to the world as we theorize about it. Via this route, physicists and mathematicians together have brought to fruition the remarkable, very mathematical gauge theories of elementary particles and fundamental interactions, which boast a very rich surplus structure and provide good evidence that, at least in their context, we cannot do physics without mathematics.

From the discussion in the second chapter we gathered that mathematics relates to physics through mappings. In our examination of this relation we discerned three different kinds of ambiguity concerning the representation of physics by mathematics. The ambiguity of the first kind, or *ambiguity of which mathematical structure to choose*, is the end result of having more than one concrete mathematical structures, which are all adequate

therefore, that the different structures we use in ambiguities of the first and of the second type have the same representational content.

However, in the third kind of ambiguity we saw that there is a conventional choice of a particular gauge from an equivalence class of gauges within the same structure, but the gauges 'live' in the surplus structure and are not mapped -at least not directly- to any physical objects whatsoever. What is more, we cannot do physics without referring to these surplus entities, hence the one-to-one correspondence between the mathematical entities and the physical objects breaks down in this case. Given that the choice of gauge seems to be conventional, the question we then asked was: *What has the conventional choice of mathematical representation of a physical system got to do with physics*? This question we will try to answer now that we have examined physical systems that are described using mathematical surplus structure, that is to say, systems with gauge symmetries.

The mathematical formalisms available to gauge theories were examined in the third chapter where we argued that at present, the best one available is that of fibre bundles. If we restricted our view of gauge theories and considered them to be constrained Hamiltonian systems, there would not be much that could be said about the relation of the surplus structure to physics. The answer to the question above, then, would have to be something pedestrian, like 'the conventional choice of the mathematical representation has got nothing to do with physics, it is just one among the many ways we could use in order to describe the system under examination'. The advantage that the unified and top-bottom fibre bundles formalism offers, on the other hand, is that the relations between the entities that live in the surplus structure and those that occur in the rest of the mathematical structure only

are expressed clearly in the form of mappings which, we believe, help us clarify the function of the surplus entities in the theory as a whole. These mappings reveal the function of the connections -or gauge fields- as information bearers and help us break free from the vicious circle of trying to attribute to them a causal character. This function of the connection is highlighted by our examination of the A-B effect and by our investigation of the possible explanations that one may give. Moreover, the purpose of the surplus structure as the descriptive tool-kit of the theory becomes manifest and help us to understand the sense in which the mathematical controls the physical .

5.1 Is Topological Explanation Justified?

The existent models of scientific explanation have been proved insufficient for several reasons, as we saw in chapter 4. Gauge theories as they stand today challenge them further because their inherently non-separable character requires holistic, rather than bit-by-bit, explanations and the existing models are not suited. This problem was elucidated when we examined the three existing attempts to provide an explanation for the A-B effect. The most promising of the three was the third one, dubbed the *C* approach by Lyre, which alleges that it is the non-vanishing loop integrals of the connections, or holonomies, around the solenoid that explain the effect. Although they did not provide a satisfactory explanation, holonomies gave a very good indication that there is more to the spacetime around the solenoid responsible for the effect than just the magnetic field which is confined inside it. The conclusion one may draw from the non-vanishing holonomies is that zero magnetic field, or zero curvature, does not imply trivial parallel transport necessarily. From the fibre

bundles theory it is known that if a region of the base or spacetime manifold is not simply connected then there appear non-trivial holonomies that describe the A-B effect in a quantitative way. There are two problems with the *C* approach. The first one is that it asserts that the non-vanishing holonomy explains the effect. To our view, if one uses Stokes theorem, one realizes that the holonomy states that somewhere within the boundaries there is some magnetic flux. But this very fact cannot explain what is happening, it only affirms some physical fact, which by the way was our starting point anyway.

The second is that it turned a sufficiency argument into a necessity one by claiming that if the holonomies are non-vanishing then the region they surround is not simply connected; but this conclusion does not follow because there may be other physical entities present, entities of which we are not aware, that are responsible for the effect. It may as well be the case that it is the nature of the electromagnetic field, which we do not really know, that is responsible for the A-B effect. In the A-B effect we get non-local results when either gauge or the electromagnetic fields are involved. Can we say from this that nature behaves in a non-local way necessarily? We don't really know, say some eminent physicists when asked[205]. The necessity they try to establish is desirable because this way we would know that the relation between the physical structure and the surplus structure is exact and that the surplus structure actually governs the physical realm. But what we can see is nothing like that. Rather, it is a consistent picture that can be used for explaining how certain physical objects (e.g. the B-field) affect the behaviour of other physical objects (e.g. the electrons) even though these objects do not interact directly with each other.

[205] In private conversations, Lee Smolin and K. Stelle have admitted this elusive necessity that does not seem to be required by nature itself.

It id for this reason that these considerations lead to the inevitable conclusion that "there is a sense in which the connection is a more fundamental object in nature than the curvature, even though a connection is gauge dependent and *not* directly measurable", as Nash and Sen put it[206], and hence to the quest for another explanation of the A-B effect.

In the physics literature, though not in the philosophical, there have been suggestions for a holistic, topological explanation of the effect, that may be explicated in two ways, as we have seen. One may claim that the topology of the base space is non-trivial because of the presence of the solenoid, a fact that results in non-vanishing holonomies that account for the effect, or one may assert that since there is a *U(1)* group acting in the structure, the topology of the vacuum is non-trivial and as a result we get the effect. We argued that in a classical context, none of these constitutes a legitimate scientific explanation because the presence of the solenoid dose not render the base space non-trivial, in the first case, while in the second the very presence of the solenoid prevents one from considering that the state of the fields is a vacuum. Hence in both cases negative analogies are contained that undermine the explanatory power of the arguments. However, when we shifted our point of view from classical to quantum relativistic we realized that there it did make sense to talk about vacuum state despite the presence of a solenoid with a magnetic field in it. At last, the topological explanation seems to work thanks to the quantum fluctuations of the relativistic vacuum. In other cases in gauge field theories where the vacuum state is global right from the start and where the solutions of the equations of motion are topological

---

206 Nash & Sen, *Topology and Geometry for Physicists*, p. 302. Bold letters in the original.

objects, the model of topological explanation which uses global topological considerations plays a genuine explanatory role, we argued.

Where does this leave topological explanations, one may ask. First of all, the suggested type of explanation certainly does not cover all possible explanations in physics, since there are plenty of examples of explanations that are not covered by it. To just mention one, take explanations in atomic physics. We saw in the previous chapter that alleged topological explanations, like the one given in the case of projectile motion in the gravitational field near the surface of the earth, do not provide any explanatory service at all. In other cases, aside from those in gauge theories, like for example in the case of 'handedness', global topological considerations that provide good explanations have been employed since the times of Kant. In *The Shape of Space* Nerlich, following Kant (1768), argues that since the property of being a left, or a right, hand cannot be a property intrinsic to hands, nor can it be some relation which they bear to other objects or to parts of space, "it must lie in a relation between hand and space as a whole, in virtue of its topology"[207] that turns out to be an aspect of its shape. If space is orientable, then the existence of incongruous counterparts, like left and right hands, is justified globally; if, however, space is non-orientable, then although locally there seem to exist incongruous counterparts, its topology does not allow for their existence globally[208]. Hence topology does a very good explanatory job in this case. Finally, in the case of gauge theories, so far as the A-B effect is concerned topological considerations in the classical case may provide only fictional, and for this reason

---

[207] Nerlich, *The Shape of Space*, p.5.

[208] Here we are not concerned with the philosophical debate about substantivalism and relationism that takes place around this issue. For details about this debate see Nerlich (1994) and Hoefer (2000).

not satisfactory, explanations but in the case of relativistic quantum field theories and of solitons topological explanations are not only legitimate but also the best we can get.

5.2 Reassessing the Relation Between Physics and Mathematics

From the perspective of fibre bundles, the connections, or gauge fields, have been attributed a different status. The challenge that all the three attempts to explain the A-B failed to meet was to attribute to the gauge potential some causal status,which, within the context of constrained Hamiltonian systems, seemed an inevitable step, especially because there seemed to be no other -obvious- way of interpreting it. By shifting our perspective and examining the effect using a different mathematical structure, we were able to actually understand the gauge field as having a different function and therefore a different status. In the fibre bundle context, the gauge potentials become the objects of the surplus structure that encode and contain all the information about the change in the topology of spacetime, that is all the information about any change in physics. the story one could tell in this context is that the connections 'communicate' to matter fields the fact that the topology is non-trivial not by causally affecting them but by 'instructing' them how to modify their phase. They do not govern the behaviour of the electrons, this is actually done by the magnetic field which constrains the choice of the gauge orbits that are allowed. Either of the possible gauges, though, can and do convey the message. Hence, if the gauge fields are given the status of information bearers, rather than causal agents, we may claim that the surplus structure is not just a superfluous mathematical artefact; rather, it contains all the information that is

necessary in order to predict the behavior of our physical system and to explain what is the cause of it -i.e. the non-trivial topology associated with the magnetic field- and how it affects it. It provides us with a quantitative method, or in other words with an entity -the connection- that predicts and describes the effects and hence enables the calculation of measurable quantities. It is the resulting non-vanishing holonomies that calculate precisely the shift in the phase of the electron, after all. The conclusion that follows from all these is that the gauge fields cannot be given the status of truly existing fields, i.e. real fields that act locally, nor can they be understood as merely objects of a purely mathematical surplus structure that has no relation to the physical system, but as objects encoding all the necessary information that is not contained in the part of the mathematical structure which is adequate for the description of the physical fields. Although the choice of a specific gauge may still seem purely conventional, the actual functional role of the gauge fields themselves in the theory goes, therefore, beyond mere convention.

The situation here is reminiscent of something that happens quite a lot in mathematics, where an extended mathematical structure describes and explains what is going on in a 'reduced' one -so much that it seems as though the extended controls the 'reduced'. Michael Redhead in his *Unseen World* (2001) discusses two such examples: the proof of Desargues' theorem in plane projective geometry and the binomial expansion of the function $\frac{1}{1+x^2}$. In the case of Desargues' theorem, in order to prove it one moves from two to three dimensions by introducing a point outside the plane; then one has only to assume the axioms of incidence to prove the theorem in the plane. As for the binomial expansion, its convergency properties are explained -or controlled, as Redhead put it- once we extend the

mathematical structure from the real numbers' line to the complex plane. A similar example from mathematics that finds application in scattering theories occurs when one tries to solve certain singular differential equations, where once again is the complex plane rather than the real number line that explains -or controls- the behavior of the functions involved. In all the cases mentioned here the surplus structure is apparently more informative and hence more powerful than the 'reduced' and the same holds in the case of gauge theories, of course.

But now let us investigate what other conclusions we may draw from this change of our perspective about the ambiguity of the third kind in gauge theories. The choice of a specific gauge in a given problem is conventional in the sense that since gauge orbits define equivalent classes, any member of an appropriate class would do. In the case of gauge theories, the question *'What has the conventional choice of mathematical representation of a physical system got to do with physics?'* can be rephrased as follows. Since the conventional choice of gauge has such an import in our understanding of the phenomena and since it is because of this possibility that we get description of interactions and, maybe, acceptable, approximate topological explanations, what can we say about the relation between the two, i.e. the relation between mathematics and physics? What we would like to claim here is that it is not the conventional choice *per se* that allows us to do so, rather it is the freedom to choose our gauge, or 'unit of measurement' in a broad sense if you prefer, that enables a complete description of what is happening or is going to happen. Allowing for freedom of choice of the available 'measuring tools' we are able to capture all the information that is needed and that is available. One should be reminded here the case of

impedances that we mentioned in chapter 2 and compare it with the case of the gauge theories. In that case the mathematical structure was also richer than the physical and with more possibilities -in the form of relations- to handle the entities involved. However, in the case of gauge theories we have a mathematical structure which, thanks to the symmetries present, is also richer in its ontology in the sense that it contains mathematical entities that do not directly correspond to the physical ones.

The fibre bundle formalism provides us with a plethora of tools and non-physical entities, or information bearers as we like to call them, which 'live' in a richer structure than that of the constrained Hamiltonian systems, or even that which we perceive as physical. This richer and hence filled-with-more-possibilities structure gives the opportunity to explain events that we are aware of using objects or descriptive tools that initially we were not aware of. Is this relation between mathematics and physics accidental? No, definitely not! So far as gauge theories are concerned, an indication that this relation is non-accidental is provided by the remarkable heuristic success of gauge theories. The discovery of all the three massive gauge fields that mediate the weak interactions, as well as of the quarks that are the messengers of the strong interactions, relied on theoretical predictions based on the 'natural' extensions of the $U(1)$ theory of electromagnetism. Of course, as the experimental data indicated discrepancies between theory and experiment, or nature, modifications of the theories followed promptly so that the disagreement ceased. One such modification was dictated by the fact that the weak gauge bosons were massive; gauge theories, on the other hand, predicted massless gauge potentials. The way out of this difficulty was provided by the so-called spontaneous symmetry breaking, which requires that a choice of gauge has

occurred such that the gauge potentials assume a fixed value and hence they acquire mass. Apart from the fact that experimental import modified the theoretical interpretation of the theory, this incident is very important for another reason. Nature indicated that in the case of weak interactions the weak interaction information bearers produced massive, measurable currents, which means that the mathematical entities corresponded to physically real particles with directly measurable properties. A possible reading of this is that the gauge fixing, which this specific kind of interactions required and which is impossible when the symmetries are still present, obliges us to move from a world of possibilities and information bearers to the world of actualities and gauge fields that correspond to physical objects.

The relation between physics and mathematics, on one hand, and physics and nature, on the other, is a dynamic relation where the choice of a particular mathematical framework for a physical theory depends on the needs and the progress of the theory on a merely theoretical but also on a phenomenological level, while often, the development of a particular branch of mathematics is also influenced by advances of some physical theory -and experiment- that made use of them. In either case, there has been an interaction between physics and mathematics. From the history of physics, the cases that exemplify this two-way relation abound. Take the startling case of general relativity, to begin with. When Einstein started on the road to this theory and looked for an appropriate mathematical framework, tensor calculus was already available for him to use. Another example where mathematics and physics developed hand in hand was Newtonian mechanics and differential calculus. But also, there are examples where mathematics developed after physics, in order to accommodate physics. One well known example is the case of quantum mechanics

and Dirac's formulation, which triggered research in mathematics that led to the development of the theory of distributions. Another example, which we have already mentioned and is perhaps less well known but very important to our case study of gauge theories, is Noether's work on variational principles and variational calculus. It was work in progress in physics and interaction with physicists who were working on that area that guided her mathematical research; of course one should not neglect the role of her intuition. Most of all, in the first chapter we discussed to some extent the history behind the genesis of gauge theories and we saw there that the mathematical framework of these theories matured not on its own but with persistent and diligent work and a lot of communication between mathematicians and physicists. But then, one may ask, what is it in this relation that makes an interaction like this possible? A key word that, to our view, is revealing of the nature of this relation is *dialectic*. The relation between physics -theoretical as well as experimental- and mathematics is a dialectic relation in which input and feedback play a crucial role.

Bibliography

Abraham, R. & Marsden, J. E. 1978. *Foundations of Mechanics*. The Benjamin/Cummings Publishing Company Inc.

Aitchison, I. J. R. & Hey, A. J. G. 1989. *Gauge Theories in Particle Physics*. Adam Hilger.

Aharonov, Y. & Bohm, D. 1959. *Physical Review.* 115: 84.

Arnol'd, V. I. & Novikov, S. P. (eds.) 1985. *Dynamical Systems IV*. Springer-Verlag.

Auyang, S. Y. 2000. Mathematics and reality: two notions of spacetime in the analytic and constructionist views of gauge field theories. *Philosophy of Science*. 67: 482-594.

———— 1995. *How Is Quantum Field Theory Possible*? Oxford University Press.

Balin, D. & Love, A. *Introduction to Gauge Field Theory*. Institute of Physics Publishing.

Belot, G. 1998. Understanding Electromagnetism. *Brit. J. Phil. Sci.* 49: 532-555.

——— 1996. *Whatever is Never and Nowhere is not: Space, Time and Ontology in Classical and Quantum Gravity*. Ph.D. Thesis, University of Pittsburgh.

Benacerraf, P. & Putnam, H. (ed.s). 1964. *Philosophy of Mathematics*. Cambridge University Press.

Bjorken, J. D. & Drell, S. D. 1965. *Relativistic Quantum Fields*. Mc Graw-Hill Book Company.

———————— 1964. *Relativistic Quantum Mechanics*. Mc Graw-Hill Book Company.

Bocchiere, P. & Soinger, A. 1978. *Nuovo Cimento 47A*: 475.

Brading, K. 2002. Which symmetry? Noether, Weyl and conservation of electric charge. *Studies in History and Philosophy of Modern Physics*. 33: 3-22

Brading, K & Brown, H. 2001. Noether's variational problem. In *Symmetries in Physics: Philosophical Reflections*. (forthcoming) Cambridge: Cambridge University Press. Brading & Castelani (ed.s)

Bremer, M. S. 1999. *Notes on D=11 Supergravity*. Unpublished.

Brown, J. R. 1999. *Philosophy of Mathematics*. London: Routledge.

Buchwald, J. Z. (ed.) 1995. *Scientific Practice*. Chicago: The University of Chicago Press.

Burgess P. J. & Rosen, G. 1997. *A Subject with no Object*. Oxford: Clarendon Press.

Cartwright, N. 1983. *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cheng & Li. 1984. *Gauge Theory of Elementary Particle Physics*. Oxford: Clarendon Press.

Chevalley, C. 1946. *Theory of Lie Groups*. Princeton: Princeton University Press.

Chihara, C. 1990. *Constructibility and Mathematical Existence*. Oxford: Clarendon Press.

Cornwell, J. F. 1997. *Group Theory in Physics, vol. II*. Academic Press.

Darling, R. W. R. 1994. *Differential Forms and Connections*. Cambridge University Press.

Dirac, P. A. M. 1964. *Lectures on Quantum Mechanics*. New York: Belfer Graduate School of Science Monograph Series.

Doughty, N. A. 1990. *Lagrangian Interaction*. Sydney: Addison-Wesley.

Drienschner, M., Eynck, T. O., Lyre, H. 2001. Comment on Redhead: the interpretation of gauge symmetry. *Ontological Aspects of Quantum Field Theories.* Khulman, Lyre & Wayne (ed.s).

Earman, J. 2000. Gauge matters. *Philosophy of Science.*

Ehresmann, C. 1943. Sur les espaces fibrés associés a une variété différentiable. *Comptes Rendus des Séances de l'Academie des Sciences.* 216: 628-630.

———— 1942. Espaces fibrés de structures comparables. *Comptes Rendus des Séances de l'Academie des Sciences.* 214: 144-147.

———— 1941. Sur les propriétés d'homotopie des espaces fibrés. *Comptes Rendus des Séances de l'Academie des Sciences.* 212: 945-950.

———— 1941. Espaces fibrés associés. *Comptes Rendus des Séances de l'Academie des Sciences*. 213: 762-764.

———— 1934. Sur la topologie de certains espaces homogènes. *Annals of Mathematics*. 35: 396-443.

Einstein, A. 1988. *The Collected Papers of A. Einstein. Vol. 8, The Berlin Years: Correspondence 1914-1918.* Princeton: Princeton University Press.

Feynman, R. P., Leighton R. B., Sands, M. 1964. *The Feynman Lectures on Physics.* Addison-Wesley Publishing Company.

Feynman, R. P. 1985. *QED The Strange Theory of Light and Matter.* Penguin.

———— 1965. *The Character of Physical Law.* Penguin.

Field, H. 1985. On conservativness and incompleteness. *The Journal of Philosophy.* 82 (5): 239-260.

——— 1980. *Science Without Numbers*, Princeton: Princeton University Press.

Fine, A. & Fine, D. 1997. Gauge theories, anomalies and global geometry: the interplay of physics and mathematics. *Studies in History and Philosophy of Modern Physics.* 28(3): 307-323.

Fleming, G. 2000. Reeh-Schlieder meets Newton-Wigner. *Philosophy of Science.* 67: 495-515.

Fonda, L. & Ghirardi, G.C. 1970. *Symmetry Principles in Quantum Physics*, New York: Marcel Dekker Inc.

Fock, V. 1927. On the invariant form of the wave and motion equations for a charged point-mass. *Zeit fur Physik.* 39: 226. (Translated in O'Raifaertaigh, 1997).

——— 1926. Zur Zur Schrödingerschen Wellenmechanik. *Zeit. fur Physik.* 36: 242-250. (Translated in O'Raifaertaigh, 1997).

Goldstein, H. 1950. *Classical Mechanics.* Addison-Wesley Publishing Company.

Grant, E. 1981. *Much Ado About Nothing.* Cambridge: Cambridge Univeristy Press.

Gribov, V. N. 1978. Quantization of non-Abelian theories. *Nuclear Physics B.* 139: 1-19.

Guillemin, V. & Sternberg, S. 1984. *Symplectic Techniques in Physics*. Cambridge University Press.

Healey, R. 2001. On the reality of gauge potentials. *Philosophy of Science*. 84 (4): 432.

——— 1999. Quntum analogies: a reply to Maudlin. *Philosophy of Science*. 66: 440-7.

——— 1997. Non-locality and the Aharonov-Bohm effect. *Philosophy of Science*. 64: 18-41.

Hesse, M. 1963. *Models and Analogies in Physics*, London: Sheed and Ward.

Hendry, J. 1984. *The Creation of Quantum Mechanics and the Bohr-Pauli Dialogue*. D. Reidel Publishing Company.

Henneaux, M. & Teitelboim, C. 1992. *Quantization of Gauge Systems*. Princeton: Princeton University Press.

Hintikka, J. 1969. *The Philosophy of Mathematics*. Oxford University Press.

Hoefer, C. 2000. Kant's hands and Earman's pions: chirality arguments for substantival space. *International Studies in the Philosophy of Science*. 14 (3): 237-255.

Huggett, N. & Weingard, R. 1994. Interpretations of quantum field theory. *Philosophy of Science*. 61: 370-388.

Isham, C. J. 1999. *Modern Differential Geometry for Physicists*. World Scientific.

Jammr, M. 1954. *Concepts of Space*. Cambridge, MA. Harvard University Press.

Kaluza, T. 1921. On the unification problem in physics. *Sitzungsber. Preuss, Akad. Wiss. Berlin*. 966.

Kant, I. 1768. On the first ground of the distinction of regions in space. Translation in Walford, D. & Meerbote, R. (1992) *The Cambridge Edition of the Works of Immanuel Kant: Theoretical Philosophy, 1755-1770*. Cambridge: Cambridge University Press.

Klein, O. 1938. Conference on New Theories in Physics, held at Kasimierz, Poland 1938. Reprinted in *1988 Conference on New Theories in Physics, Proc. XI Warsaw Symposium on Elementary Particle Physics*. Ajduk, Z., Pokorski, S., Trautman, A. (eds.).

——— 1926. Quantum theory and five-dimensional relativity. *Zeit fur Physik*. 37: 895. (Translated in O'Raifaertaigh, 1997).

Kobayashi, S. & Nomizou, K. 1969. *Foundations of Differential Geometry, vol. II.* Interscience Publishers.

———————— 1963. *Foundations of Differential Geometry, vol. I.* Interscience Publishers.

Koertge, N. 1984. Galileo and the problem of accidents. *Journal of the History of Ideas.* pp.389-408.

Köhler, W. 1942. *Die physichen Gestalten in Ruhe und in stationären Zustand.* Braunschweig.

Koperski, J. 2001. Has chaos been explained? *Brit. J. Phil. Sci.* 52: 683-700.

Leeds, S. 1999. Gauges: Aharonov, Bohm, Yang, Healey. *Philosophy of Science. 66*: 607-627.

Lewis, D. 1986. *In the Plurality of Worlds.* Oxford: Basil Blackwell.

Liu, C. 2001. Infinite systems in SM explanations: thermodynamic limit, renormalization (semi-) groups and irreversibility. *Philosophy of Science*, 68 (*Proceedings*): S325-S344.

London, F. 1927. Quantum-mechanical interpretation of Weyl's theory. *Zeit. fur Physik.* 42: 375. (Translated in O'Raifaertaigh, 1997).

Lyre, H. 2001a. A versus B! Topological non-separability and the Aharonov-Bohm effect. *Contribution for the International IQSA Conference: Quantum Structures V.* Cesena/Cesenatico, Italy.

——— 2001b. The principles of gauging. *Philosophy of Science.* 68: 371-381.

——— 2001c. Comment on Redhead: the interpretation of gauge symmetry. *Ontological Aspects of Quantum Field Theories.* Khulman, Lyre & Wayne (ed.s).

——— 2000. A generalized equivalence principle. arXiv:gr-qc/0004054

——— 1999. Gauges, holes and their 'connections'. *Lecture at Fifth International Conference on the History and Foundations of General Relativity, 1999, University of Notre Dame.* Notre Dame. Indiana. gr-qc/9904036

Malament, D. 1982. Review of Science Without Numbers. *The Journal of Philosophy.* 79 (9): 523-534.

Maudlin, T. 1998. Discussion: Healey and Aharonov-Bohm. *Philosophy of Science.* 65. 361-368.

Miller, D. 1994. *Critical Rationalism.* Chicago: Open Court.

Mills, R. & Yang, C. N. 1954. Isotopic spin conservation and a generalized gauge invariance. *Physics Review.* 95: 631.

——————— 1954. Conservation of isotopic and gauge invariance. *Physics Review.* 96: 191.

Nash, C. & Sen, S. 1983. *Topology and Geometry for Physicists.* Academic Press.

Nagel, E. 1979 (1st ed.1961). *The Structure of Science.* Hackett Publishing Company.

Nakahara, M. 1990. *Geometry, Topology and Physics.* Institute of Physics Publishing Ltd.

Nerlich, G. 1994. *The Shape of Space.* Cambridge: Cambridge University Press.

Newton-Smith, W. H. (ed.) 2000. *A Companion to the Philosophy of Science.* Blackwell.

O'Raifeartaigh, L.1997. *The Dawning of Gauge Theories*, Princeton: Princeton Series in Physics.

Pais, A. 1986. *Inward bound.* Oxford: Clarendon Press.

Pauli, W. 1953. Meson-Nucleon Interaction. Letters to A. Pais.

Peshkin, M. & Tonomura, A. 1989. *The Aharonov-Bohm Effect.* Springer-Veralg.

Pokorski, S. 1987. *Gauge Field Theories.* Cambridge: Cambridge University Press.

Putnam, H. 1967. Mathematics without foundations. *The Journal of Philosophy.* 64(1): 5-22

Quine, W. V. 1970. *Philosophy of Logic.* Prentice Hall.

——————— 1966. *The Ways of Paradox and other Essays.* New York: Random House.

Redhead, M. 2002. The interpretation of gauge symmetry. *Ontological Aspects of Quantum Field Theories.* Khulman, Lyre & Wayne (ed.s).

———— 2001. The intelligibility of the universe. In *Philosophy in the New Millenium.* O'Hear, A. (ed.)

———— 2001. *The Unseen World.* LSE Series.

———— 1999. Review of S. Y. Auyang: How is Quantum Field Theory Possible? *British Journal for the Philosophy of Science.*

———— 1995a. *From Physics to Metaphysics.* Cambridge: Cambridge University Press.

———— 1995b. More ado about nothing. *Foundations of Physics.* 4: 1443-7.

———— 1995c. The vacuum in relativistic quantum field theory. Hull, Forbes & Burian (ed.s). *PSA 1994, vol.2.* East Lansing: Philosophy of Science Association. 77-87.

Resnik, M. D. 1997. *Mathematics as a Science of Patterns.* Oxford: Clarendon Press.

Ruben, D. H. (ed.) 1993. *Explanation.* Oxford: Oxford University Press.

———— 1990. *Explaining Explanation. London*: Routledge.

Russell, B. 1927. *The Analysis of Matter.* London: Allen & Unwin.

———— 1919. *Introduction to Mathematical Philosophy.* London: Allen & Unwin; rep. New York: Dover.

Ryckman, T. 2001. Weyl's debt to Husserl. In *Symmetries in Physics: New Reflections.* (forthcoming) Cambridge: Cambridge University Press. Brading & Castelani (ed.s)

Ryder, L. H. 1985. *Quantum Field Theory.* Cambridge: Cambridge University Press.

Salmon, W. C. 1989. *Four Decades of Scientific Explanation.* Minneapolis: University of Minneapolis Press.

———— 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Schrödinger, E. 1926. Quantization as an eigenvalue problem. *Annalen der Physik* 81: 162. (Translated in O'Raifeartaigh 1997.)

———— 1922. On a remarkable property of the quantum orbits of a single electron. *Zeit. f. Physik* 12: 13 (Translated in O'Raifeartaigh 1997.)

Schweber, S. S. 1994. *QED and the Men Who Made It: Dyson, Feynman, Schwinger and Tomonaga*. Princeton: Princeton University Press.

Shanks, N. (ed.) 1998. *Idealization IX: Idealization in Contemporary Physics*. Rodopi.

Shapere, A. & Wilczeck, F. 1989. *Geometric Phases in Physics*. Singapore: World Scientific.

Shapiro, S. 2000. *Thinking about Mathematics*. Oxford: Oxford University Press.

———— 1983. Conservativeness and incompleteness. *The Journal of Philosophy* 80 (9): 521-531.

Shaw, R. 1955. *Ph.D. Thesis*. University of Cambridge.

Shenker, O. 1994. Fractal geometry is not the geometry of nature. *Stud. Hist. Phil. Sci.*, 52(6): 967-981.

Singer, I. M. 1978. Some remarks on the Gribov ambiguity. *Commun. Math. Phys*. 60: 7-12.

Smolin, L. 1997. *The Life of the Cosmos*. Oxford: Oxford University Press.

Smith, P. 1998. *Explaining Chaos.* Cambridge: Cambridge University Press.

Scholz, E. 1994. Hermann Weyl's contribution to geometry, 1917-1923. *The Intersection of History and Mathematics*. Chikara, Mitsuo, Dauben (ed.s). Birkhäuser Verlag.

Steenrod, N. 1951. *The Topology of Fibre Bundles*. Princeton: Princeton University Press.

Tung, W. K. 1985. *Group Theory in Physics*. Singapore: World Scientific.

Utiyama, R. 1956. Invariant theoretical interpretation of interaction. Phys. Rev. 101 (5): 1597-1607.

van Fraassen, B. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.

———————— 1980. *The Scientific Image*. Oxford: Oxford University Press.

Wald, R. M. 1984. *General Relativity*. Chicago: The University of Chicago Press.

Weinberg, S. 2000. *The Quantum Theory of Fields*, vol. III. Cambridge: Cambridge University Press.

———— 1996. *The Quantum Theory of Fields*, vol.I & II. Cambridge: Cambridge University Press.

———— 1993. *Dreams of a Final Theory.* London: Hutchinson Radius.

———— 1972. *Gravitation and Cosmology.* New York: Wiley.

Weyl, H. 1950. A remark on the coupling of gravitation and electron. *Physical Review.* 77(5): 699-701.

——— 1929: Electron and gravitation. *Zeit. fur Physik.* 330: 56.

——— 1918. Gravitation and electricity. *Sitzungsber. Preuss, Akad. Berlin.* 465. (Translated in O'Raifaertaigh, 1997).

Whitney, H. 1940. On the theory of sphere-bundles. *Proc. Nat. Ac. Sci.* 26: 145-153.

———— 1937. Topological properties of differentiable manifolds. *Proc. Nat. Ac. Sci.* 43: 785-805.

———— 1935. Sphere-Spaces. *Proc. Nat. Ac. Sci.* 21: 464-468.

Wigner, E. 1967. *Symmetries and Reflections.* Bloomington: Indiana University Press.

Wu, T. T. & Yang, C. N. 1975. Concept of non-integrable phase factors and global formulation of gauge fields. *Physical Review D.* 12 (12) : 3845-3857.

Yang, C. N. 1974. Integral formalism of gauge fields. *Phys. Rev. Let. 33* (7): 445-447.