Models of the World, Data-models and the Practice of Science: The Semantics of Quantum Theory

Mauricio Suárez
London School of Economics and Political Science

Thesis submitted for the degree of
Doctor of Philosophy of London University

August 1997

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

Abstract

The most important problems in the philosophy of quantum mechanics are the problem of measurement and the problem of the 'acausal' EPR correlations. It is commonly thought that these problems call for a new *interpretation* of the quantum theory. I argue that it is possible to construe both problems rather differently, as resulting from a mistaken understanding of scientific theory-application. It then becomes possible to tackle both problems independently of questions of interpretation, by attending carefully to what constitutes a successful application of a scientific theory, and of the quantum theory in particular.

In the first part of the Thesis, I argue against a standard conception of scientific theory-application. This standard conception, which is often presupposed in the philosophical discussions of quantum mechanics, takes the applications of a scientific theory to constitute its domain of empirical adequacy. I argue that, on the contrary, a scientific theory can be applied to phenomena that it does not subsume. I present a case study in the history of superconductivity to illustrate and to motivate this claim.

In the second part, I argue that the problem of measurement can be construed as the impossibility of applying the quantum theory to measurement interactions. I then argue that Arthur Fine's proposed solution to the measurement problem implicitly abandons the standard conception of application. Finally I look at quantum correlation phenomena. Bas Van Fraassen has claimed that the EPR correlations fit no causal model. The correctness of this claim depends on what probabilistic constraints a causal model is taken to have to satisfy. I argue, following Nancy Cartwright, that Van Fraassen's constraints on common-cause models are too strong; and I describe a direct-cause model that, I urge, constitutes a successful application of the quantum theory to the EPR correlations.

Contents

0.1 Introduction

It is often assumed that the domain of empirical adequacy of a scientific theory is constituted by all of the theory's applications. This assumption has its simplicity in its favour, but it also occassionally yields some seemingly counterintuitive results. Quantum theory, for instance, would be in a dramatically much better shape than the general theory of relativity: the applications of quantum theory are many, diverse and widely accessible; those of general relativity are few, and highly remote. Thus, quantum theory should be *much better* confirmed than general relativity –and yet, scientists seem generally to think that both theories are just about as equally likely –or as equally unlikely– to be true.

In part I of the Thesis, I argue for a separation between application and confirmation: a scientific theory does not gain confirmation from all of its applications. In many instances of successful theory-application, a mediating model is confirmed instead. In these cases a scientific theory will play a largely instrumental role in its own application. The theory, although playing an essential part in the process of generating an accurate model of the phenomena, will not itself be *under test* in that process. This seems to point towards a distinction between epistemic and pragmatic reasons to uphold a theory. I outline a contrast between degree of 'confirmation' and degree of 'confidence' that, I suggest, may be capable of tracking this distinction. Confirmation, on this view, accrues only via those applications where the theory plays a genuinely propositional role.

But the common assumption is not totally off the mark. For indeed, success in applying a theory must count for something; it must somehow raise our *confidence* in the theory. I argue that while a theory's degree of confirmation is an indication of its truth, or of its empirical adequacy, a theory's degree of confidence measures its reliability as an instrument in application.

From this point of view, quantum theory looks very much like any other physical theory. It can be applied with the intention of testing it, as is currently being done by Leggett and his collaborators, who are investigat-

ing its validity in the macroscopic domain; or it can be simply applied, but not tested –as is the case in so many of its technological applications. In the second part of the Thesis, I argue that some of the foundational problems surrounding quantum mechanics appear in a very different light when construed as problems of application rather than, as often understood, as problems of empirical adequacy, and confirmation.

A word is due on my usage of the terms 'realism' and 'instrumentalism'. In the first part of the Thesis, I argue against a particular form of scientific realist epistemology, embodied in what I call the 'idealization account' of application. The idealization account involves a very strong form of scientific realism about physical theory: just by investigating the theory it ought to be possible to derive all of its possible descriptions of the physical objects and problem-situations in its domain. In some of his writings Ernan McMullin defends this form of realism; but it is clear to me that scientific realists are not generally required to adopt such a strong view. There are many other forms of scientific realism that I do not discuss in this Thesis, and against which the arguments in part I have no force. Ian Hacking's entity realism, Nancy Cartwright's causal realism and Jim Brown's phenomenological realism are but a few examples.

In the second part of the Thesis, I also employ the terms 'realism' and 'instrumentalism' –but in a different fashion, as *methodological*, rather than epistemic, attitudes. My view is that, in bringing out the instrumental reliability of a scientific theory, either an instrumentalist or a realist methodology may prove fruitful. In chapter 3 I exhibit a case where, I think, an instrumentalist approach helps; while in chapter 4 I show how a more realistic attitude can also prove useful. Thus, about methodological issues I take an undogmatic attitude: any number of strategies may help advance the aim of instrumental reliability. Both instrumentalism and realism can play a part in raising confidence in our best scientific theories.

0.2 Acknowledgements

It would be impossible to acknowledge everyone who has helped with this project over the last five years. The following, however, stand out. I am extremely grateful to Nancy Cartwright for her patience and dedication as my supervisor. This Thesis no doubt owes much to her book *How the Laws of Physics Lie* and to discussions with Nancy –surely in content, but especially in inspiration. I also want to thank Margaret Morrison for discussing her work with me in such great detail. Martin Jones has been my main transatlantic 'point of reference': APA and PSA meetings would have been a lot less fun without him. At the LSE, Thomas Uebel taught me about the logical empiricists; Moshe Machover and John Worrall taught me logic; and Elie Zahar taught me general relativity. Among the students, Towfic Shomar helped with superconductivity and Marco DelSeta with quantum measurement, and LaTex.

Over the years Michael Redhead and Jeremy Butterfield have regularly welcomed me to their seminar meetings in Cambridge, where I have met many of their students. During 1995/96 I replaced Harvey Brown as lecturer in philosophy of physics at Oxford; that was my first proper academic post, and I know how lucky I was to have Harvey as my first philosophical colleague. Some other people who helped at Oxford include Paul Castell, Michel Ghins, Rom Harré and Simon Saunders. During 1996/7 I was a Teaching and Research Fellow in the Department of Logic and Metaphysics, St Andrews: I thank everyone there, particularly Peter Clark, Stephen Read and Crispin Wright, for their support.

My parents have been a source of relentless support and encouragement. My sister Alejandra, an economist, shares many of my concerns and anxieties about knowledge, method, and practice. My friends Sarah, Danny, Cherry, Chris, Crispin, Hamish and Lucy encouraged me in many different ways. This Thesis is dedicated to Chrissie –she deserves a very special thanks for her support; for sharing all my frustrations –and my joys; and, more than anything, for gently reminding me that there are other very good things in life besides philosophy.

Further Acknowledgements

A shortened version of chapter 1 will appear in a Cambridge University Press volume entitled *Models in Physics and Economics* [102]. Aspects of the case study in section 1.4 have been published in a joint paper with Nancy Cartwright and Towfic Shomar ([34] (but see footnote 16). Section 3.2. on the modal interpretation, is a version of a paper published in *Foundations of Physics Letters* [123].

Later drafts of the case study in superconductivity in chapter 1 developed in friendly competition, if not outright distant cooperation, with Steven French and James Ladyman. I had a very encouraging, albeit brief, email correspondence with Kostas Gavroglu when I was beginning work on the case study in superconductivity; the subsequent publication of his biography of Fritz London was timely, and provided further encouragement. My interest in the modal interpretation was started while attending the *IV Symposium on Foundations of Physics*, in Köln in June 1993. A version of section 3.3 was read at the *II Symposium in Fundamental Problems in Quantum Physics*, Oviedo, July 1996. And a version of chapter 2 was read at the *Einstein Meets Magritte* conference in Brussels in May 1994. I want to thank the organizers of those three conferences, – Peter Mittelstaedt and Paul Busch, Miguel Ferrero, and Diderik Aerts–, for enabling me to attend by awarding me generous grants. I also thank Adolf Grünbaum for inviting me to attend his lectures at the Symposium in Philosophy and Cosmology in Santander, September 1995.

Others who have helped include Sven Aerts, David Albert, Wolfgang Balzer, Joseph Berkovitz, David Bloor, Otavio Bueno, Jordi Cat, Gian-Piero Cattaneo, Hasok Chang, Rob Clifton, James Cushing, Marisa Dalla Chiara, William Demopoulos, Dennis Dieks, Andrew Elby, Arthur Fine, Mathias Frisch, Roberto Giuntini, Stephan Hartmann, Meir Hemmo, Robin Hendry, Colin Howson, R.I.G. Hughes, Paul Humphreys, Matthias Kaiser, Ed Jurkowitz, Federico Laudisa, Christoph Lehrner, Nicholas Maxwell, Ernan McMullin, Mary Morgan, Philip Percival, David Papineau, Stathis Psillos, Steven Shapin, Bas Van Fraassen, Jos Uffink, Andrew Warwick.

I also thank my various coauthors: Fred Müller and Pieter Vermaas [106], Nancy Cartwright and Towfic Shomar [34], Marco DelSeta [43], and Harvey Brown and Guido Bacciagaluppi [19]. Although in this Thesis I do not discuss any joint work, I am in no doubt their collaboration has influenced my understanding of the issues that I do discuss.

Word Count: 61,000 words approx.

Part I

Scientific Application and Empirical Adequacy

Chapter 1

Mediating Models, Superconductivity and Realism

1.1 Models as Mediators

There are many kinds of models in science. In the second chapter I shall review and discuss some of them. In this chapter I focus on a specific kind: mediating models. First, in this section, I introduce the notion of a mediating model, and I briefly outline some of its main features. In the remaining sections I make the notion more precise by considering the key role that mediating models play in the application of scientific theories, and the implications of mediating models for the epistemology of science.

Mediating models have been recently discussed by a number of authors. Adam Morton[1] has referred to them as the providers of physical insight; Margaret Morrison[2] has studied and discussed their properties carefully;

[1] 'Mathematical Models: Questions of Trustworthiness' [105]; also in conversation, Bristol May 1997.

[2] 'Mediating Models: Between Physics and the Physical World' [104].

among historians of science, Norton Wise[3] has unearthed some of the mediating models and instruments that operated in Enlightment France; and a forthcoming book[4] collects recent work by philosophers of physics and economics that deals with methodological and epistemological issues arising from this type of models.

1.1.1 Features of Mediating Models

Mediating models always stand between theory and the physical world. Their main function is to enable us to apply scientific theory to natural phenomena. A mediating model often involves a novel conception of a particular physical phenomenon that facilitates the application of some established physical theory to such phenomena. Morrison has identified three main features. First, mediating models are not derivable from theory. In a very specific sense the construction of these models is not *theory-driven*; I will emphasise this feature later on in this chapter. Second, these models are not necessitated by the empirical data either (although they can be suggested by the phenomena). In contrast to a data-model which is determined by the data together with established statistical techniques, a mediating model '*is more than simply a phenomenological classification constructed as a convenient way of representing [data]*' (Morrison [104, page 11]). In other words, mediating models typically involve substantial theoretical and conceptual assumptions. Finally mediating models have a very significant property: they can replace physical systems as the central objects of scientific inquiry. Morrison [104, page 9] writes:

> Not only do models function in their own right by providing solutions to and explanations of particular problems and processes, but in some cases they even supplant the physical system they were designed to represent and become the primary object

[3] 'Mediations: Enlightenment Balancing Acts, or the Technologies of Rationalism' [141].

[4] *Models in Physics and Economics*, Cambridge University Press, Cambridge [102].

> of inquiry. In other words, investigation proceeds on the basis of the model and its structural constraints rather than the model being developed piecemeal in response to empirical data or phenomena.

This is an essential feature of mediating models; it distinguishes this type of models from other closely related types, such as for instance Heinz Post's floating models. As reported by Michael Redhead[5], floating models will also satisfy the first two features ascribed to mediating models. Redhead [114, page 158] describes a floating model as:

> ... a model which is disconnected from a fundamental theory T by a computation gap in the sense that we cannot justify mathematically the validity of the approximations being made but which also fails to match experiment with its own (model) predictions. So it is disconnected from the fundamental theory and the empirical facts. In Post's graphic terminology the model 'floats' at both ends. It has, in this sense, no theoretical or empirical support.

Post's parody of a floating model was an example he called the Farm Gate Contraction. Redhead [114, page 158] reports this example as follows:

> A farmer investigates the relation between the length of the diagonal strut and the length of the rails and stiles of a farm gate. Although he is familiar with Euclid the derivation of Pythagoras's theorem is utterly beyond his deductive powers. So he invents a model theory, a linear one, in which the lengths are related by $l = x + y$ instead of $l = \sqrt{x^2 + y^2}$. Now [the model] has many properties analogous to [the theory] –for $x = 0$ or $y = 0$ it gives correct values for l and l increases monotonically with x or y in the model as in the correct theory. But detailed

[5] Redhead, 'Models in Physics' [114].

> measurement shows that [the model] is false. So the farmer now introduces a new effect, the Farm Gate Contraction, to explain the mismatch between the predictions of the model and the experimental results.

The Farm Gate Contraction is a correction on a floating model. The model, even when corrected in this way, is certainly not *required* by the data, as is shown by the fact that there are alternative models that fit the data just as well (the 'correct' theory is one of them); and it is not supported by any fundamental theory as it is only an inspired (although ultimately mistaken) initial guess. Floating models are not derivable from either theory or empirical data. In that sense a mediating model is a kind of floating model.

However a mediating model has a further essential feature, one that is not necessary for a floating model. While a floating model may convey no new knowledge at all, a mediating model mediates between high level theory and the world by conveying some *particular* or *local* knowledge specific to the effect or phenomenon that is being modelled. This is why the model itself becomes the active focus of scientific research. While a floating model is typically only a computational tool, a mediating model is a carrier of specific, or 'local' knowledge. Morrison [104, page 12] writes:

> It is exactly in these kinds of cases, where the model takes on a life of its own, that its true role as a mediator becomes apparent. Because investigation centres on the model rather than nature itself its representative role is enhanced to the point where the model serves as a source of mediated knowledge rather than as simply a mediator between high level theory and the world.

Hence this third feature, the capacity a model may have to replace the phenomenon itself as the focus of scientific research, is an essential feature of mediating models. It distinguishes mediating models from the far larger class of floating models. In this chapter I develop a further feature of mediating

models, which is essential for a full understanding of the role that these models play in the application of scientific theories. Mediating models will often fix the criteria that we use to refine our theoretical descriptions of a phenomenon. These criteria are required to apply theory successfully to the world. Before discussing this fourth feature of mediating models it may be worth emphasising the differences with some of the the types of model I shall be discussing in the second chapter.

1.1.2 Mediating Models in the Philosophy of Science

The philosophical lessons to be learnt from this new role of models as mediators are perhaps still unclear. Some preliminary remarks may serve to illustrate why it would seem that there *must* be profound implications. The syntactic view of scientific theories equates models with interpretations of theory. This tradition assimilates the distinction between scientific theories and scientific models to the syntax/semantics distinction in linguistics. The theory is a purely syntactical entity, while the models provide us with the semantics of the scientific discourse. The relation between the models and the theory is one of satisfaction: the model must make the theory's axioms true.

It is difficult to see how models are to literally 'mediate between theory and the world' if the view of models as providing the semantics of theories is correct. If models are interpretations, or partial interpretations, of theories they are in a sense supererogatory on theory. A theory will define an elementary class of models; hence it will greatly restrict the class of permitted models. An inconsistent theory, for instance, restricts the class of permitted models to the empty set. However, it is a presupposition of the notion of models as mediators that there are three distinct objects (theories, models, and the world) and that they are ordered with the theory at the most abstract end, the world at the opposite end, and the model as the interface between the two. Moreover the model conveys specific physical knowledge. The view of models as interpretations of theories allows for a trichotomy between theory, model and world but it seems to order these objects the

wrong way around, with models at the most abstract end, and theories at the interface (as model/theory/world rather than as theory/model/world). Moreover it implies that models do not convey any significant novel physical information that is not already encoded in theories. Surely this is partly the reason why proponents of this view have so often attempted to construe the relation of confirmation as a purely syntactical connection between a theory, on the one hand, and evidence, on the other.

It is possible on the syntactic view to see the world itself as a possible model of a theory. The theory is a set of axioms in some formal system, and it implicitly defines an elementary class of models. We may then say that a theory is true if it has the world as one of its models, and false if the world is not among its models. In so far as the world itself is to be a model, the distinction between model and the world collapses, and we are left with a dichotomy theory/world. So on this view, models mediate between the theory and the world only in the sense that the set of permitted models of a theory can be said to include the world itself. The activity of model-building reduces, on this account, to investigating ways the world would have to be if some specific scientific theory was true. This assumes, once more, that the totality of scientific knowledge about the world is encoded in theories.

There is also, of course, the semantic conception of theories that I shall describe in the second chapter, advocated by Suppes, Van Fraassen and others. Here the distinction between theory and model collapses as, according to the semantic view, theories *are* models –they are really nothing but collections of models. On this view there is a hierarchical structure of models, from low-level data-models to high-level theoretical models. So the contrast between theories and models disappears. I shall address the semantic view of theories in the second chapter, where I describe the conception of empirical adequacy within the semantic view.

1.2 The Idealization Account of Application

In this section I describe a specific proposal for theory-application that involves models as idealizations. This proposal, essentially due to Ernan McMullin, is intended to go further than the traditional accounts of scientific theorising, by placing the activity of model-building at the very core of scientific practice. I argue, in section 3, that despite its intention, McMullin's proposal effectively dispenses with the need for models as mediators because it invariably construes models as approximations to theories. In section 4 I try to illuminate and explicate this practical role of models as mediators by using an example from the history of superconductivity. In section 5 I discuss the epistemological implications.

1.2.1 Forms of Idealization

How does scientific theory get applied to the world? Ernan McMullin[6] has proposed a realist account of theory-application. Theoretical descriptions, argues McMullin, are always idealized; they apply only under very special circumstances, often not realizable in practice. But the idealization inherent in theory is not epistemologically problematic. Although theoretical descriptions are often not *absolutely* true or false, they are *approximately* true or false.

McMullin finds support for this view in Galileo's idealization techniques. In *The New Sciences* Salviati, Galileo's alter ego, argues against the Aristotelian views of some of Galileo's contemporaries, personified mainly in the character of Simplicio. The discussion centres around the techniques of approximation required to apply theory to concrete problem situations and to validate the theoretical claims of Galilean mechanics. Two examples are repeatedly used: parabolic trajectories of projectiles, and motion of rolling objects on inclined planes. Consider the latter. Galileo's claim is of course that the motion of a perfectly symmetric sphere under the earth's

[6] McMullin, 'Galilean Idealization' [100].

gravitational pull on a frictionless plane in a vacuum follows a very strict mechanical law. But any real plane will exhibit friction, any real object is bound to be only imperfectly spherical, and in any actual experiment there is bound to be dampening due to the presence of air. To establish his mechanical conclusions on the basis of actual experiments, Galileo has to claim that the imperfections can be accounted for, and that there is a well established and unique method of introducing corrections into theory to account for 'impediments', the imperfections of nature.

In order to show that there is indeed such a method, Galileo (and McMullin) need to appeal to the notion of approximation. There are, broadly speaking, two methods for approximating theory to the world. One is the approximation of the theory to the problem situation brought about by introducing corrections into the theoretical description (the theory is refined to bring it closer to the problem-situation). The other is the approximation of the problem-situation to the theory by means of simplifications of the problem-situation itself. In the latter case the theory is left untouched, while the problem-situation is altered; in the former case the converse is true: the problem-situation is left untouched, while the theoretical description is corrected.

Let us first consider the former kind of approximation whereby the theoretical description is refined to bring it closer to the problem-situation. This is a form of approximation towards the real case: the corrections introduced into the theoretical description are intended to account for the imperfections that occur in the problem-situation. The same method can be reversed (n.b. this is not yet the second method of approximation) by *subtracting*, rather than adding, the required corrections. We may call this an *idealization*; for the result of such subtraction is of course a more, rather than less, idealized description of the problem-situation. The important feature of this idealization is that the subtraction of corrections is performed on the theoretical construction, while the description of the problem-situation is left entirely unaffected. For this reason McMullin[7] calls the first form of approximation

[7]See McMullin [100, page 256].

construct idealization.

The second method of approximation brings the problem-situation closer to theory. We idealize the description of the problem-situation, while leaving the theoretical construction unaffected. McMullin calls this *causal idealization* because the description of the causes present in the problem-situation is altered to bring the description into the domain of the theory. In the practice of physics this process can come in either of two forms. It can come first in the form of conceptual redescriptions of the problem-situation, performed only in thought, and not in reality. In such 'thought-experiments' interfering causes are idealized away and the result is a simplified description of the problem-situation. Secondly, there is also the possibility of physical 'shielding' of the experimental apparatus, which will involve changes in the actual experimental set-up. Such changes are designed to minimise the influence of interfering causes, or to block such influences out altogether. It is perhaps instructive to quote Galileo in full:

> We are trying to investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistance, so that the whole difference of speed existing between these moveables would have to be referred to inequality of weight alone. Hence just one space entirely void of air –and of every other body, however thin and yielding– would be suitable for showing us sensibly that which we seek. Since we lack such a space, let us (instead) observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant. If we find in fact that moveables of different weight differ less and less in speed as they are situated in more and more yielding media, and that finally, despite extreme difference of weight, their diversity of speed in the most tenuous medium of all (though not void) is found to be very small and almost unobservable, then it seems to me that we may believe, by a highly probable guess, that in the void all speeds would be entirely equal. (quoted in McMullin [100, page 267])

It is uncertain whether Galileo performed any of these experiments in actual fact. If he did, he would certainly have needed to use a technique of 'shielding' to minimise the influence of interfering causes. If, on the other hand, he didn't actually perform the experiments then in this passage he is describing a series of thought-experiments that gradually minimise the effects of interfering causes –in the mind, of course, not in reality. The dynamics of moveables in the void that he concludes will exhibit equal speeds is in either case a *causal* idealization. Starting with a concrete problem-situation (i.e. the motion of an object in the earth's atmosphere) Galileo constructs a set of simpler problem-situations. If relations between quantities measurable in these gradually simpler thought experiments converge to a law we can then enunciate the law for the ideal (simplest) case. The resulting law is a *causal* idealization, because the simplifications correspond to missing causes in the problem-situation.

McMullin summarises the main features of each form of idealization concisely as follows:

> We have seen that idealization in this context takes on two main forms. In construct idealization, the models on which theoretical understanding is built are deliberately fashioned so as to leave aside part of the complexity of the concrete order. In causal idealization the physical world itself is consciously simplified; an artificial ('experimental') context is constructed within which questions about law-like correlations between physical variables can be unambiguously answered. Causal idealization, instead of being carried out experimentally, can also be performed in thought, when we focus on the single causal line in abstraction from others and ask 'what would happen if'. (op. cit. [100, page 273])

In this Thesis I focus only on *construct idealization*, the kind of idealization whereby simplifications are worked out on the theoretical description, rather than on the problem situation. This is partly because I believe that

every case of theory-application will involve, in practice, at least some degree of construct idealization; and partly because this is the only form of idealization that shall concern us when the main issues in the philosophy of quantum mechanics are addressed in Chapters 3 and 4 of the Thesis. Let me just stress that *construct* idealization requires no thought-experiments, nor does it require tampering with the real experimental situation. Only one problem-situation, namely the real case, is entertained. It is the theoretical description that gets modified by introducing correction factors that represent 'impediments', the special circumstances that make up the particular problem-situation. In other words, in construct idealization, the theoretical description is refined gradually to make it applicable to the problem-situation.

In actual practice we look for approximations to the theory that can be applied to a particular problem-situation. Michael Redhead[8] refers to these approximations as *impoverishment* models. The theoretical description may be very complicated: there may be no analytic solutions to the theoretical equations. How then can we derive the correct impoverishment model? How can we choose among all possible approximations the very one that accurately represents the behaviour of the system? The important point, that I shall now stress, is that the theory itself must contain the information required to select the correct approximation if the approximation in question is to count as a *de-idealization* of theory.

1.2.2 Idealization and Scientific Realism

A theory can be applied by finding a simplifying approximation to it that is adequate for the description of a phenomenon. Not all approximations, however, guarantee that the theory is confirmed by its applications. It is essential to McMullin's realism that the corrections introduced into the theoretical description should not be *ad hoc*. The corrections have to be well motivated *from the point of view of theory.* If the theory is to receive confir-

[8]Redhead [114].

mation boosts from its applications the corrections need to be not only consistent with the theory, but also if not dictated by, at least *suggested by*, the theory. If in a particular application the necessary corrections turned out to be inconsistent with the theory, the theory could be said to be disconfirmed; if the corrections were consistent with the theory, but not suggested by it, the theory would neither receive a confirmatory boost nor a disconfirmatory one. McMullin explicitly acknowledges this important point: according to the (*construct*) idealization picture of theory application the manipulations exerted on the theoretical description must be 'theory-driven' because the theory itself is to be truth-apt (a 'candidate for truth' in Ian Hacking's terminology [9]) and is to gain confirmation through its applications. If the corrections were not suggested by the theory then the resulting description would be *ad hoc* and, from the point of view of a realist epistemology, it would be unable to provide any evidence for the truth of the theory. Thus McMullin writes:

> The implications of construct idealization, both formal and material, are thus truth-bearing in a very strong sense. Theoretical laws [...] give an approximate fit with empirical laws reporting on observation. It is precisely this lack of perfect fit that sets in motion the processes of self-correction and imaginative extension described above [i.e. *deidealization*]. If the model is a good one, these processes are not *ad hoc*; they are suggested by the model itself. Where the processes *are* of an *ad hoc* sort, the implication is that the model is not a good one; the uncorrected laws derived from it could then be described as 'false' or defective, even if they do give an approximate fit with empirical laws. The reason is that the model from which they derive lacks the means for self-correction which is the best testimony of its truth. (op. cit [100, page 264]).

[9] See Hacking [79].

In this passage McMullin is not using the term 'model' to describe a mediating model, as I do in this chapter. I have taken 'mediating models' to be distinct from established theory while McMullin is here taking 'model' to stand for a theoretical description, as in the semantic view of theories. McMullin makes it clear that the corrections introduced into a theory to generate predictions in a particular physical problem-situation have to be suggested by the theory itself; otherwise, the corrections would be *ad hoc* and the resulting description, no matter how well it fitted the particular case, would not yield any confirmatory boost for the theory. If the corrections were not suggested by the theory there would be no way to account for the effects that those corrections have upon the final predictions. As McMullin notes [100, page 256] it is essential that there be "*a way of dealing with the fact that construct idealizations "depart from the truth". If this departure is appreciably large, perhaps its effect [...] can be estimated and allowed for*". By requiring that the corrections into a theoretical model be well motivated from the point of view of theory we make sure that we are always able to estimate their contribution to the final description.

In other words, application must be *theory-driven* in order to provide confirmation for the theory. I shall refer to this sort of theory-driven approximation of the theory to the problem-situation that results in a refinement of the theoretical description as *construct de-idealization*, or *deidealization* for short. The name is due to the fact that an approximation of this kind is nothing but the converse process to construct idealization. In forming construct idealizations we idealize away, by subtracting from the description, those features of the problem-situation that are either (a) irrelevant to the theoretical description, or (b) relevant to the theoretical description, but are also known in the theoretical description to have effects that are precisely accountable for. (In the latter case construct idealization is often used for the purpose of improving the mathematical tractability of the problem.) In either (a) or (b) a strict criterion of theoretical relevance is presupposed. It is the theory that tells us what are the relevant features to be idealized away, and suggests how to account for their effects. The same criterion of theo-

retical relevance must be in place if the converse process of "adding back" features is to count as a meaningful *deidealization.* The requirement that the introduction of corrections into a theoretical model be well motivated from the point of view of theory ensures that the criterion is firmly in place.

The above discussion is perhaps sufficient to make clear why the idealization account of theory application satisfies the realist's constraints. For applications which follow the idealization account, the theory receives confirmation boosts from the applications. The corrections that serve to generate successful applications are necessarily consistent with theory, because they are suggested by theory. They are corrections suggested by some strict relevance criterion –a criterion that is wholly and unambiguously theoretically determined. So, an application of a theory that conforms to nature provides a good reason to believe that the theory itself is true.

Let me now briefly address the sense of 'approximate truth' that is involved in the idealization account. McMullin is not arguing that scientific theories are approximately true or false. The theory, on McMullin's view, contains its own criteria of application; so, indeed, the theory contains all of its possible theoretical descriptions of the problem-situation. Hence the theory is either true (if it contains one true description of the problem-situation), or false (if it contains none). It is because of this that a successful deidealization of a scientific theory to a particular problem-situation should always be taken as an indication of the theory's truth: it shows that the theory contains one true description of the problem-situation.

McMullin's claim is rather that *theoretical descriptions* of a particular problem-situation may be approximately true or false. His intuition is roughly as follows: successive approximations of a theory to a problem-situation have a degree of confirmation inversely proportional to their 'distance' from the problem-situation as measured on the 'idealization scale'; but, –for a realist–, degree of confirmation is degree of truth; so 'distance in the idealization scale' measures degree of truth. Given two representations A and B of some concrete problem-situation if A is less idealized than B then, in a very precise sense, A is *truer* than B. To pursue a Galilean exam-

ple: the representation of a sphere rolling down a frictionless plane is less idealized if described in the actual atmosphere (description A) than if described in a vacuum (description B). The description in the atmosphere has to involve a measure of the dampening due to air. The realist would want to claim that this description is *truer* than the description of the sphere in a vacuum, in a totally unobjectionable sense of the notion of objective truth. For a scientific realist, such as McMullin, Galilean idealization provides the *model* for the notion of approximate truth.

1.3 The Role of Models in Theory-Application

It is always open to the opponent of realism to attack the inference from the past success of a theory to its future success, and from its pervasiveness in practice to its truth. An instrumentalist may after all have no qualms with Galilean idealization: it is a technique of application, it is often used, and sometimes with some conviction that it carries epistemic weight, but in fact it is only a tool, and it can give no genuine warrant for belief other than the psychological comfort offered by the familiarity of its use. But here I do not attempt a general philosophical rebuttal of the realist view. This would take us one step back, in the direction of the traditional disputes concerning arguments for scientific realism –disputes that have not been settled, possibly because they could never be settled[10].

On independent grounds the realist view won't work. The realist wants to claim that the idealization account captures the essential features of the procedure of theory-application. I argue in this chapter that the idealization account is seriously flawed because it can not explain the role of models in scientific practice. The inadequacy of the idealization account stems from the fact that, in practice, theory-application does not typically follow the pattern of *deidealization*. But the realist does not rest content with this base-

[10]In the philosophy of science this *quietism*, or perhaps simply 'pessimism', towards the realism/antirealism debate has been most ably defended by Arthur Fine –see chapters 7 and 8 of *The Shaky Game* [62].

level claim; in addition he claims that the idealization account also agrees with scientific practice at an *epistemological level.* Scientists' confidence in a scientific theory typically increases on account of its many successful applications. The realist seeks support for the idealization account also on these epistemological practices of scientists. And, indeed, on the idealization account a theory gains confirmation through its applications.

To sum up, there are two distinct claims that the realist makes on behalf of the idealization account. First, this account agrees with the practice of theory-application and second, it agrees with scientific epistemology. In this Thesis I contest the truth of the former claim, and I argue that the latter claim, although true, does not provide ammunition for the realist account of theory-application.

1.3.1 Problems with Idealization

I like to illustrate the idealization account of application with a simple example in mechanics due to Ronald Giere[11]. The example brings out very clearly what, in my view, is the major defect of the idealization account.

Consider the derivation of the equation of the damped linear oscillator from that of the simple harmonic oscillator. The equation of the simple harmonic oscillator is:

$$m\frac{d^2x}{dt^2} = -\left(\frac{mg}{l}\right)x, \tag{1.3.1}$$

while the equation that describes a damped harmonic oscillator is:

$$m\frac{d^2x}{dt^2} = -\left(\frac{mg}{l}\right)x + bv. \tag{1.3.2}$$

The process that takes one from the theoretical description of the frictionless harmonic oscillator to the damped harmonic oscillator is a successful deidealization in the attempt to apply classical mechanics to a real-life pendulum. The extra term bv represents the dampening due to air friction that

[11] Giere [73, chapter 3].

any real oscillator must be subject to. The introduction of this correction term into the idealized description afforded by the equation of the simple harmonic oscillator is motivated by theoretical considerations: in classical mechanics friction is modelled by a linear function of velocity[12]. By introducing well-motivated corrections into the theoretical description of the simple harmonic oscillator we obtain a less idealized description of a real-life pendulum in ordinary circumstances, namely the description of a damped harmonic oscillator.

Equation (1.3.2) tends to equation (1.3.1) in the limit $b \longrightarrow 0$, as required for an approximation. Hence the two descriptions agree in the asymptotic limit. There are of course plenty of equations that, just like (1.3.2), tend to the original equation (1.3.1) in some mathematical limit. (1.3.2) is special because it is derived from the equation of the simple harmonic oscillator by a process of deidealization. The damped harmonic oscillator and the simple harmonic oscillator are objects defined implicitly in the theory by their satisfaction of the corresponding equations; hence it is the theory that determines the relations between them. The correction terms introduced into the equation of the simple harmonic oscillator are justified by the putative relations between the objects themselves. Equation (1.3.1) is satisfied by a linear oscillator with no friction; equation (1.3.2) is satisfied by a linear oscillator subject to friction. The theory contains all the necessary techniques to represent this difference formally.

Hence the idealization account makes superfluous the use of models in theory application. Theories must be seen as entirely self-sufficient in the task of generating genuinely realistic representations of problem-situations[13]. Where the idealization account is true, or generally true, it follows that models cannot *mediate* between theories and the world: in the application

[12]For a discussion of modelling friction see eg. Herbert Goldstein [76, page 24]

[13]Specifically, and to anticipate the main issue in what is to follow, theories do not (*must not*) rely on independently-standing models in order to fix the corrections required for successful *deidealizations*.

of scientific theories that satisfy the idealization account, there is essentially no work for mediating models to do.

This is perhaps a bit too hasty, as it does seem intuitively possible to supplement the idealization account so as to accommodate the role of models as mediators. A defender of the idealization account could argue that we are misusing the terms "theory" and "model". She could argue that these terms refer to relative, rather than absolute, concepts. Consider two theoretical descriptions A and B, and suppose that A is less idealized than B with respect to some physical system S. A could then be defined as "a model with respect to" B, and B as "a theory with respect to" A. The proponent of the idealization account can now claim that the role of models as mediators is a necessary consequence of these relative definitions. Given the idealization account "models" always mediate between theory and the world because a "model" is defined to be a relatively deidealized theoretical representation of a problem-situation in the domain of the "theory".

Although this approach in terms of relative definitions is certainly possible, it does not seem plausible for at least two reasons. First, the relative definition of models and theories seems counterintuitive *precisely* when it is conjoined with the idealization account. We already saw that the latter account entails that the less idealized a description of a situation, the *truer* it is of that situation. The relative definition entails that less idealized descriptions are "models" with respect to the more idealized ones: it follows that "models" are generally *truer* than "theories". Such conclusion runs against the common view that theories are candidates for truth or falsehood while models, as fictional representations, are not truth-apt. This is obviously not a conclusive objection because the intuition that underlies the common view can always be resisted, but it makes the relative approach less plausible.

Another reason against the relative definitions is that intuitively the definitions of "model" and "theory" should be transitive. Given three representations, A, B and C, if C is a theory with respect to B and B is a theory with respect to A then C should also be a theory with respect to A. There is no reason, however, why the relation of idealization should be

transitive, and hence no reason why the proposed relative definitions should turn out to exhibit transitivity either. In other words, it may well be the case that A is less idealized than B, and B is less idealized than C, and yet A is not less idealized than C. On the relative definition approach, C would be a theory with respect to B and a model with respect to A, even if B is a theory with respect to A. The possibility of failure of transitivity is a result of the fact that the relation "to be less idealized a representation than" is really a three-place relation, not a two-place one, depending as it does crucially upon the problem-situation for which the representations are intended. When assessing which of two representations is more idealized it is assumed that they are intended representations of the same problem-situation. But of course a representation may be intended of several problem-situations. The failure of transitivity described above could arise for instance if B was such a representation, intended for two different problem-situations.

In practice we will often be concerned with classes of theoretical descriptions that won't allow for any kind of ranking. For instance, when two applications of the same theory to two different problem-situations are considered, there is no way to rank the two applications. It does not make sense to say of any one of them that it is the *more idealized*: they are incommensurable as regards their degree of idealization. Similarly there is no way to establish which of two applications of two different theories to the same problem-situation is *more idealized.* In both cases the relative definition in terms of degree of idealization entails that it is impossible to assess which one is a model and which a theory.

Perhaps these two problems can be surmounted. The proponent of the idealization account may convince us that our intuitions about the absolute character of the definition of "models" and "theories" need to be corrected, and that the relative definitions are sound after all. There would still be a third problem for the idealization account which in my view cannot be surmounted. The idealization account assumes there is a final representation of every system in the theory's domain of application. In practice we may never be able to write this representation, as it may be hideously compli-

cated; but the representation must exist because it can be approximated to an arbitrary degree by successive deidealizations of the theory.

However, even in the simple case of the harmonic oscillator the presumption that such a final theoretical representation exists seems profoundly perplexing. The equation of the damped harmonic oscillator is certainly not a final representation of this kind. It is not a theoretical representation of any concrete real system in the world. Admittedly the equation of the damped harmonic oscillator is a less idealized representation than the equation of the simple harmonic oscillator for real-life penduli. But this does guarantee that the theory contains a (true) representation of a real-life pendulum. The theory may be incomplete; there may well be some aspects of the problem-situation left unaccounted for, even after all the relevant corrections suggested by the theory have been added in.

But now the promised sense in which models were to mediate between theory and the world is definitely lost: models mediate only between theory and further models. On the idealization account the theory does all the work required for its own application by determining, in stages, sets of increasingly *less idealized* representations. These representations, however, may never truly represent anything real at all.

1.3.2 The Problem of Material Abstraction

Nancy Cartwright has argued[14] that the Galilean method of idealization is valid, *if* (and only if) a metaphysics of capacities is presupposed. Cartwright is at pains to distinguish *idealization* from *abstraction*, as she wants to argue that where capacities are involved abstract laws –rather than idealized laws– hold. She draws the distinction as follows:

> ... in idealization we start with a concrete object and we mentally rearrange some of its inconvenient features –some of its

[14] In chapter 5, –entitled 'Abstract and Concrete'–, of her book *Nature's Capacities and Their Measurement* [27]. Also in 'Capacities and Abstractions' [30], and in 'How we Relate Theory to Observation' [31].

> specific properties– before we try to write down a law for it. The paradigm is the frictionless plane. We start with a particular plane, or a whole class of planes. Since we are using these planes to study the inertial properties of matter, we ignore the small perturbations produced by friction. But in fact we cannot just delete factors. Instead we replace them by others which are easier to think about, or with which it is easier to calculate. The model may leave out some features altogether which do not matter to the motion, like the colour of the ball. But it must say something, albeit something idealizing, about all the factors which are relevant.
>
> [...] By contrast, when we try to formulate [an abstract law], we consider the causal factors out of context all together. It is not a matter of *changing* any particular features or properties, but rather of *subtracting*, not only the concrete circumstances but even the material in which the cause is embedded and all that follows from that. (*Nature's Capacities and Their Measurement* [27, page 187]

In idealization we make assumptions that we know to be false about a particular system. For instance we may imagine a plane with a perfectly polished surface, even if we know well that no real surface can be *perfectly* polished: all surfaces are rough and grainy to some degree. By contrast, in an abstraction we simply *delete* a feature, and we do not substitute it with anything else. Following the same example, an abstraction would be a 'surfaceless' plane. Of course our ordinary concept of a 'plane' essentially involves the feature of its having a surface; it is part of what we mean by 'plane' that it has a surface. So a 'surfaceless plane' is really no plane at all, not even an idealized one. Unlike an idealization, which continues to be *about* real objects, an abstraction is no more *about* real planes, or real rolling balls, but is strictly about those features that remain after the subtraction takes place.

It is perhaps easiest to bring out the contrast by considering the distinction between idealized and abstract laws. An idealized law is essentially a ceteris paribus law. It is true or false of the phenomena but it holds only when a large number of antecedent idealizing conditions are satisfied. ('*If* the surface of the plane was perfectly polished, and a perfectly polished ball was to roll down the plane, and the experiment was performed in a vacuum, and... then the motion of the ball would be precisely described by Galileo's dynamic laws'.) By contrast, in an abstract law we drop the ceteris paribus conditions altogether. An abstract law is not intended to be true or false of the phenomenal world (although it may be true or false of the powers, tendencies or capacities that underlie the phenomena), and hence the satisfaction of the long idealizing antecedent is not required for the law to apply. It follows that the language of approximate truth, which is intuitive when dealing with idealized laws, is useless when it comes to abstract laws. These laws are not intended to be true or false of the phenomenal world, so they can hardly be 'approximately' true descriptions of the phenomena. As Cartwright puts it:

> [...] where relevant features have been genuinely subtracted, it makes no sense to talk about the departure of the remaining law from truth, about whether this departure is small or not, or about how to calculate it. These questions, which are so important when treating of idealizations, are nonsense when it comes to abstractions. (op.cit. [27, page 188])

Although the conceptual difference between an idealized law and an abstract law is very important, from the purely formal point of view there may be no apparent differences, as the antecedent conditions may be elliptical. Consider the equation of the simple harmonic oscillator, equation (1.3.1): does it correspond to an idealization or an abstraction? Earlier on I argued, following McMullin, that the simple harmonic oscillator is an idealization, relative to the problem-situation of real-life penduli. I assumed that there would be a long list of conditions, stating states of affairs in the world, that would need to be satisfied if (1.3.1) is to precisely describe the motion of

any real object at all. When we move to the law that contains the equation of the damped harmonic oscillator (equation (1.3.2)) we are effectively deleting one of the conditions in the long antecedent (namely the condition that reads: '*if* there was no damping due to air friction...'). The resulting law is *less idealized* because it has fewer antecedent conditions; it is, as it were, one step closer to the truth.

However, the simple harmonic oscillator can also function as an abstraction. In such case equation (1.3.1) should not be said to 'approximately describe' the motion of a real object (such as a real pulsating pendulum) but rather to *exactly* describe the motion of an abstract one (an object implicitly defined by equation (1.3.1)). Hence in abstraction we ignore the long list of antecedent conditions and we assert that the abstract law is true or false, not of the phenomenal world, but of an abstract object. Ronald Giere[15] defends this view in his book *Explaining Science*:

> I propose that we regard the simple harmonic oscillator and the like as *abstract entities* having all and only the properties ascribed to them in the standard texts. The distinguishing feature of the simple harmonic oscillator, for example, is that it satisfies the force law $F = -kx$. The simple harmonic oscillator, then, is a constructed entity.

I shall not discuss Giere's views in any detail here. Rather I concentrate on the question: can this proposal in terms of abstraction help to patch up the realist account of application?

Nancy Cartwright has argued that the realist account of application, in terms of the introduction of corrections into a representation, is a fine method for what she calls *concretization*, the converse of abstraction. Formally the process of *concretization* looks like *deidealization*. Both require the gradual introduction of corrections into a representation, as in the case of the derivation of the equation of the damped harmonic oscillator from that

[15] Giere, [73, page 78].

of the simple harmonic oscillator. However, in deidealization the refinements aim to an increasingly more *realistic* representation of a phenomenon or an entity in the world (such as, for instance, a real life pendulum). In concretization, by contrast, the corrections are intended to represent the *effects* of additional features that figure in a more concrete representation of the phenomenon, or its underlying causal structure, or a more concrete object (perhaps defined implicitly by its corresponding equation, in this case by equation (1.3.2)). But there is no guarantee that the movement is towards the *real* case: the distinction abstract/concrete does not need to parallel the distinction ideal/real.

Moreover the refinement that corresponds to a concretization need not provide evidence for a scientific theory. A scientific theory will typically not have the resources to provide enough corrections to generate a description of all the relevant features of most concrete objects to give a totally accurate prediction of the behaviour of the object in question. This is what Cartwright calls the *problem of material abstraction.* She puts it in the language of laws:

> The abstract law is one which subtracts all but the features of interest. To get back to the concrete laws that constitute its phenomenal content, the omitted factors must be added in again. But where do these omitted factors come from? [...] given a theory, the factors come *from a list.* But the list provided by any given theory, or even by all of our theories put together, will never go far enough. There will always be further factors to consider which are peculiar to the individual case. I call this the 'problem of *material abstraction* [...] After a while, it seems, in any process of concretization, theoretical corrections run out and the process must be carried on case by case. (op.cit. [27, pages 207 and 209]).

She gives several examples of this problem. Here is one:

> Donald Glaser built the first bubble chambers, using diethyl ether. He operated on the principle that a passing charged particle has the capacity to cause bubbling in a liquid in a superheated state. (The liquid is ready to boil and just needs a stimulant.) He was also successful with hydrogen and most of the heavy liquid hydrocarbons, like propane or freon. But surprisingly the bubble chamber did not work with xenon. Here the passing charged particles excite optical transitions, and the energy is sent off as light rather than producing the heat necessary to trigger the boiling. Again mundane facts about actual materials and their construction made a difference, facts outside the domain of the initial theory about the behaviour of superheated fluids. (op.cit. [27, page 209])

Cartwright's problem of material abstraction is that in any genuine concretization of a theory to a *particular* object, the criteria that we follow to choose the appropriate corrections will vary from case to case. The problem that I am setting for idealization is indeed a very similar one, but with one significant difference: it does not need to appeal to the peculiarities of any one particular concrete object. For, in fact, the list of corrections provided by the theory sometimes runs out even earlier than Cartwright's problem of material abstraction suggests. Even in the development of a totally general electromagnetic theory of superconductivity –one that, for instance, pays no attention to the peculiarities of any one specific superconducting material–, the theory on its own is unable to determine the corrections required for its application. The theory needs help from an independently motivated mediating model.

1.4 How Models Mediate: The Case of Superconductivity

The inference from successful application to the truth of the theory becomes problematic when it is noticed that in practice the criteria of theoretical rel-

evance presupposed by the idealization account are rarely operative in cases of successful theory-application. On the contrary it is often the case that scientific representations of effects or phenomena are not arrived at as deidealizations of theory. My case study in superconductivity illustrates one way in which models typically mediate between theory and the world[16]. The first theoretical representation of the Meissner effect was not found by applying a criterion of theoretical relevance for the introduction of corrections into the electromagnetic equations of a superconductor. These correction terms were not given by, and could not have been given by, classical electromagnetic theory but were rather derived from a new *model* of superconductivity. The model was motivated directly by the phenomena, not by theory. The criterion required for the application of electromagnetic theory could only be laid out when the model was in place, and an adequate classical electromagnetic description of superconductivity (the London equations) could then finally be derived.

This is, I want to claim, an important sense in which models *mediate*: they establish the corrections that need be introduced into a theory in order to generate many of its applications. My case study shows how the derivation of a theoretical representation of a physical effect can result from corrections that are suggested by a mediating model, which is independent from theory. The approximation used to generate an appropriate representation is not a deidealization of theory, because the criterion of relevance that guides the introduction of corrections is not theoretically motivated.

I have chosen the Londons' account of superconductivity for a number of reasons: first, because it is such a well-known episode of successful theory-application; second, because of the high esteem and reputation of the two scientists involved; finally, because it is a case of application that is to a large extent explicitly not a deidealization. My case study is not exceptional or isolated; on the contrary, I believe that it is paradigmatic of the activity of

[16] Aspects of this case study have been published in a joint paper with Nancy Cartwright and Towfic Shomar [34]. In this Thesis I make use only of those parts that were strictly written by me, based entirely upon my own research.

theory-application in many branches of physics.

1.4.1 The Hallmarks of Superconductivity

The electromagnetic treatment that Fritz and Heinz London[17] proposed for superconductors in 1934 is one of the most celebrated cases of theory application in the history of twentieth century physics. It was the first comprehensive electromagnetic theory of superconductivity and it remained the fundamental account of superconductivity for nearly 20 years until the advent of the BCS theory (which was heavily informed by the Londons' account, as were all subsequent theories of superconductivity). Superconductors are materials that exhibit extraordinary conducting behaviour under specific circumstances. The hallmarks of superconducting behaviour are the following two well established phenomenological findings: resistanceless conductivity and the Meissner effect.

In 1911 Kamerlingh Onnes[18] found that when mercury is cooled below $4.2K^{o}$ its electrical resistance falls to near zero. In 1914 he discovered that the effect does not take place in the presence of an intense magnetic field. This is the first phenomenological trait of superconductivity: under a certain critical transition temperature, and in the absence of strong magnetic fields, a superconductor exhibits almost perfect resistanceless conductivity. Almost perfect resistanceless conductivity is confirmed by the presence of a stationary current through, say, the surface of a superconducting ring. The current flows at virtually the same constant rate and does not die off.

The relation between the transition temperature (T_C) and the critical magnetic field (B_C) was explored experimentally by Onnes himself. He found that the following relation held with an accuracy of a few percent:

$$B_C = B_O\{1 - \left(\frac{T}{T_c}\right)^2\}. \tag{1.4.3}$$

[17] London and London, 'The Electromagnetic equations of the Supraconductor' [98].

[18] Onnes [110].

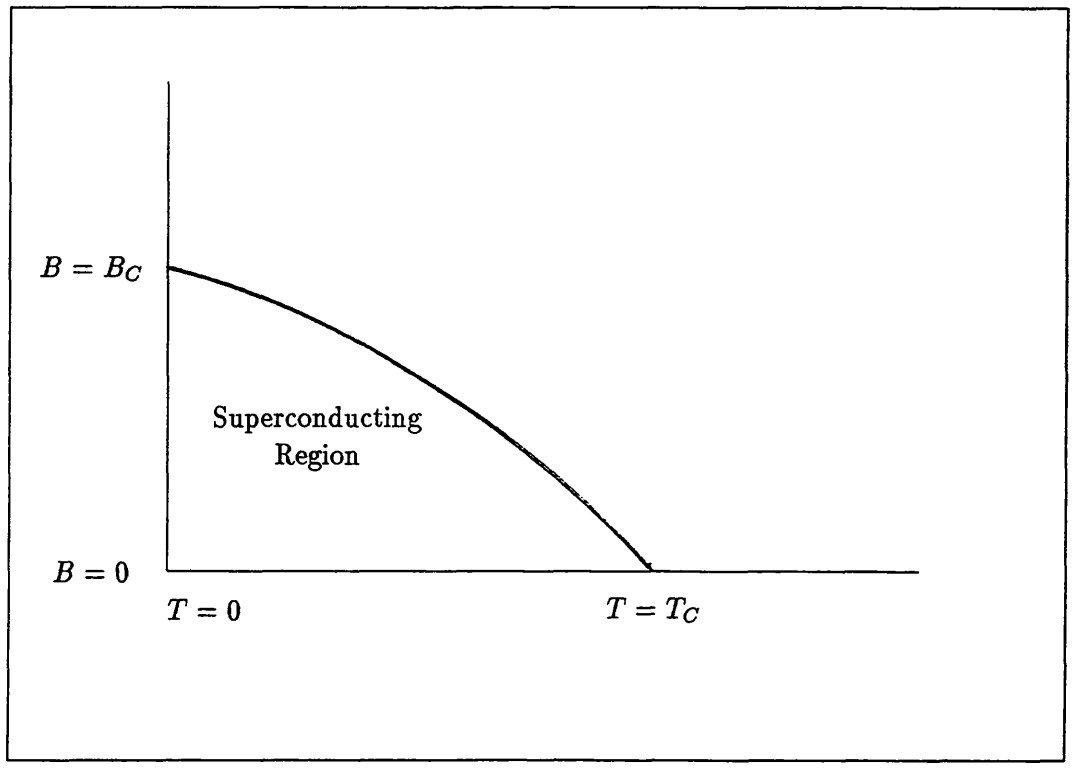


Figure 1.1: The Domain of Superconductivity

This equation defines the domain of superconductivity. Figure 1.1 is the graph of equation (1.4.3); it displays the region where superconductivity occurs. B_0 is a numerical constant which depends on the substance. The graph clearly shows that there are two different ways to approach the superconducting regime. One way is to bring down the ambient temperature to T_C while maintaining constant the weak external magnetic field (weaker than B_C). The other way is to decrease the magnetic field below B_C and to maintain constant the temperature at some $T < T_C$. Both strategies will work. Superconducting behaviour is suddenly exhibited when the critical phase transition takes place.

The second, equally important, trait of superconductivity was found in 1933 by Meissner and Ochsenfeld[19]. The *Meissner effect* is the sudden expulsion of magnetic flux from a superconductor when cooled below its transition temperature. The flux in a superconductor is always vanishingly small, regardless of what the flux inside the material was immediately before the phase transition into the domain of superconductivity took place[20].

1.4.2 Applying Electromagnetism

Superconductivity was initially considered an electromagnetic phenomenon and providing an electromagnetic treatment became the main theoretical task. This was a formidable task in view of the Meissner effect. Maxwell's equations on their own are totally ineffective: for a medium of perfect conductivity (a '*super*conductor') Maxwell's equations are inconsistent with the Meissner effect. Perfect conductivity occurs when the scattering of electrons in a medium of low resistance is so small that the electric current persists even in the absence of a supporting external electric field. For a conductor in a vanishingly small electric field, for which $\mathbf{E} = 0$, Maxwell's second equation $\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}$ predicts that $\frac{\partial \mathbf{B}}{\partial t} = 0$ and hence that $\mathbf{B}$, the magnetic field, must remain constant in time in the transition to the superconducting state. In other words, Maxwell's equations predict that the flux through a coil surrounding the metal must remain unaltered during the phase transition. The experiments of Meissner and Ochsenfeld showed that in fact there is a sudden change in the value of the external magnetic field, consistent with the total expulsion of the magnetic flux density from within the

[19] Meissner and Ochsenfeld [101].

[20] A distinction is usually made between Type I and Type II superconductors. In Type I superconductors *all* magnetic flux is expelled in the phase transition. In Type II superconductors the expulsion is only partial. Type II superconductors only appeared much later, and the distinction played no role in the historical instance that I wish to discuss. In this Thesis by "superconductors" I refer to type I superconductors only. These are thin films made out from metals like zinc, aluminium, mercury, lead.

superconductor[21].

Of course by 1933 there was much more to electromagnetic theory than just Maxwell's equations. In the construction of their theory of perfect conductivity Becker, Sauter and Heller[22] had to appeal to further assumptions about the media, the shape of the conductor, the forces that propelled the electrons in the absence of electric fields and, crucially, the form of the law that linked the electric current to external fields. Their 'acceleration' theory accounted for a persistent current in a superconductor, but it was shown by the Londons to be in contradiction with the Meissner effect.

In a normal conductor the current either induces an external electric field or is supported by one, and Ohm's law predicts that the current is directly proportional to the field, $\mathbf{j} = \alpha\mathbf{E}$. With the discovery of resistanceless conductivity Ohm's law had to be abandoned for superconductivity because the current persists in the absence of an external field. Nevertheless all proposed treatments of superconductivity continued to assume that there existed some relation between the superconducting current and external electric fields – not a proportionality relation obviously, but *some* relation nevertheless. The Londons' fundamental contribution was to make unambiguously clear that superconducting currents are in no way supported by electric fields, but by magnetic fields.

What prompted the Londons' suggestion? Why did previous attempts to understand superconductivity continue to assume that the current was physically linked to electric fields? The answer cannot be found by inspecting the state of electromagnetic theory in 1933. No significant contribution or substantive addition to the theory was made during these years that could help to explain the Londons' breakthrough. The significant event was the proposal, by the Londons, of a new *model.*

[21]The inconsistency of the Meissner effect, perfect conductivity with $\mathbf{E} = 0$, and Maxwell's equations is often emphasised in textbook discussions (see, for instances, Bleaney and Bleaney [13, chapter 13] and H.E.Hall [81, chapter 11]).

[22]Becker, Sauter and Heller [11].

Historically, the discovery of the Meissner effect signalled the turning point. This unexpected discovery brought about a change in the *conception* of superconductivity. A superconductor was initially conceived in analogy with ferromagnetism: just as a ferromagnet exhibits a magnetic dipole moment in the absence of any supporting magnetic fields, a superconductor exhibits a permanent current even if unsupported by electric fields. The superconducting current is constant in the absence of an electric field, and what this indicates is that the field is not proportional to the current, as in Ohm's law. As a replacement Becker, Sauter and Heller proposed the following 'acceleration equation'. where the field is proportional to the time derivative of the current:

$$\Lambda \frac{d\mathbf{j}}{\mathbf{dt}} = \mathbf{E} \tag{1.4.4}$$

where $\Lambda = \frac{m}{ne^2}$ (a constant that depends upon the mass m, charge e and number density of electrons n). In the absence of an external field ($\mathbf{E} = 0$) the 'acceleration equation' predicts a permanent current: $\frac{d\mathbf{j}}{dt} = 0$.

1.4.3 Enter the Model

The Londons understood that the Meissner effect pointed to an entirely different model. They modelled a superconductor as one huge diamagnet, and replaced Ohm's law with a new electromagnetic relation between the superconducting current and the *magnetic* field. The Londons went on to attempt a microscopic explanation of the coherence of the 'magnetic dipoles' in terms of a coherent macroscopic quantum superposition[23].

By modelling a superconductor as a diamagnet the Londons were able to introduce an important correction into the 'acceleration equation' theory of Becker, Sauter and Heller. Diamagnetism is associated with the tendency of

[23]Superconductivity is of course ultimately a quantum phenomenon. The definitive quantum treatment was given in 1951 by Bardeen, Cooper and Schrieffer [10] who explained the emergence of coherence by appealing to the formation of Cooper pairs at low temperatures. The history of the BCS theory is fascinating in its own right, but it is of no relevance to my present argument.

electrical charges to shield the interior of a body from an applied magnetic field[24]. Following a proposal by Gorter and Casimir[25], the Londons began by assuming that a real superconductor is constituted by two different substances: the normal and the superconducting current. They then proposed that Ohm's law be restricted to the normal current in the material, and the description of the superconducting current be supplemented with an equation that determined the relation of the current to the background magnetic flux. The 'London equation' for the superconducting current takes the form:

$$\nabla \times \Lambda \mathbf{j} = -\frac{1}{c}\mathbf{H} \tag{1.4.5}$$

where $\mathbf{j}$ is the current, and $\mathbf{H}$ represents the magnetic flux inside the superconductor.

It is important to understand that this equation was not derived from electromagnetic theory, but was suggested by the new model of diamagnetism. Although analogy was certainly involved, this is not just simply a case of reasoning by analogy. The Meissner effect does not just mean that the equations that describe magnetic flux in a superconducting material must be formally analogous to the equations for flux in a diamagnetic material. It rather means that a superconductor *is* a kind of diamagnet. Equation (1.4.5) was derived from a correction to the solutions of the old 'acceleration equation' theory -a correction prompted by the conception of the superconductor as a diamagnet. According to this conception the fundamental property of a superconductor is not nearly perfect conductivity but, of course, the expulsion of the magnetic flux within the material during the transition phase. Superconductivity is no longer characterised as the limit of perfect conductivity, but as the limit of perfect diamagnetism. Hence the phenomenon of the expulsion of the magnetic flux cannot, and should not, be explained by the emergence of a superconducting current. Super-

[24]See, for instance, Charles Kittel, *Introduction to Solid State Physics* [94, Chapter 14].

[25]Gorter and Casimir, [78].

conductivity is truly characterised by two independent and non-reducible phenomenological hallmarks: perfect conductivity and the Meissner effect.

In the theory of Becker, Sauter and Heller the absence of an electric field entails that the Meissner effect is impossible, as expected from our initial consideration of Maxwell's second equation in the case of perfect conductivity. Indeed the 'acceleration equation' entails the following equation for the magnetic flux inside the superconductor:

$$\Lambda c^2 \nabla^2 \frac{d\mathbf{H}}{dt} = \frac{d\mathbf{H}}{dt}. \tag{1.4.6}$$

Integrating with respect to time one finds the following nonhomogeneous equation:

$$\Lambda c^2 \nabla^2 (\mathbf{H} - \mathbf{H_0}) = \mathbf{H} - \mathbf{H_0} \tag{1.4.7}$$

$\mathbf{H_0}$ denotes the magnetic field at the time t=0 (ie. before the transition phase has occurred). Its value depends entirely on the value of the ambient field because a superconductor behaves exactly like a normal conductor before the phase transition, and the external field penetrates completely. The solutions to this equation are given by $\mathbf{H} = \mathbf{e}^{\sqrt{\Lambda}cx} + \mathbf{H_0}$, where the exponentials $e^{\sqrt{\Lambda}cx}$ decrease very quickly with distance x from the surface of the material. So the 'acceleration equation' predicts that the field inside a superconductor will remain invariant throughout the phase transition. No change in the external flux will be observed and a surrounding coil will experience null induction. As London and London[26] write of the theory of Becker, Sauter and Heller:

> The general solution means, therefore, that practically the original field persists for ever in the supraconductor. The field $\vec{H}_0$ is to be regarded as 'frozen in' and represents a permanent memory of the field which existed when the metal was last cooled below the transition temperature [...] Until recently the existence

[26] London and London [98, page 72].

> of 'frozen in' magnetic fields in supraconductors was believed to be proved theoretically and experimentally. By Meissner's experiment, however, it has been shown that this point of view cannot be maintained.

On the other hand the Londons' diamagnetic model suggests that the field inside the material once the transition has occurred decreases very quickly with distance x from the surface of the material. So the correct solutions must exclude the value ($\mathbf{H_0}$) of the initial field, and must contain only the exponentials $e^{\sqrt{\Lambda}cx}$. These are solutions to the following homogenous equation: $\Lambda c^2 \nabla^2 \mathbf{H} = \mathbf{H}$. From this equation, the fundamental equation of superconductivity (1.4.5) can be derived, since $\nabla \times \mathbf{H} = \frac{1}{c}\mathbf{j}$.

A useful way to visualize this result is by plotting the paths that would be followed in the graph for the phase transition (see the graph in figure 1.2 –which is taken from a standard textbook[27]). Suppose that we begin in the position designated by P in the graph. There is no initial field in the superconductor and hence no flux. In that case the prediction of the 'acceleration equation' model is as in the Meissner experiment: after the transition to the superconducting domain we expect to end up in the position given by S. In the final state there is no flux inside the bulk of the superconductor. But consider the case when *there is* initially some flux in the superconductor. For instance suppose that before the transition takes place the material is placed in some magnetic field. As the material is perfectly conducting the field will penetrate totally, and there will be a magnetic flux in the material identical to the flux outside the material. We can represent such a situation by position Q in the graph. In this case the prediction of the 'acceleration equation' theory contradicts the Meissner effect. According to the 'acceleration equation' after the transition we end up in R as the flux is 'frozen in' the material; the Meissner effect entails that we end up in S nonetheless.

To sum up, the Londons suggested that the superconducting current is maintained by a magnetic field. The relation is of inverse proportionality, so

[27] Bleaney and Bleaney [13, page 399].

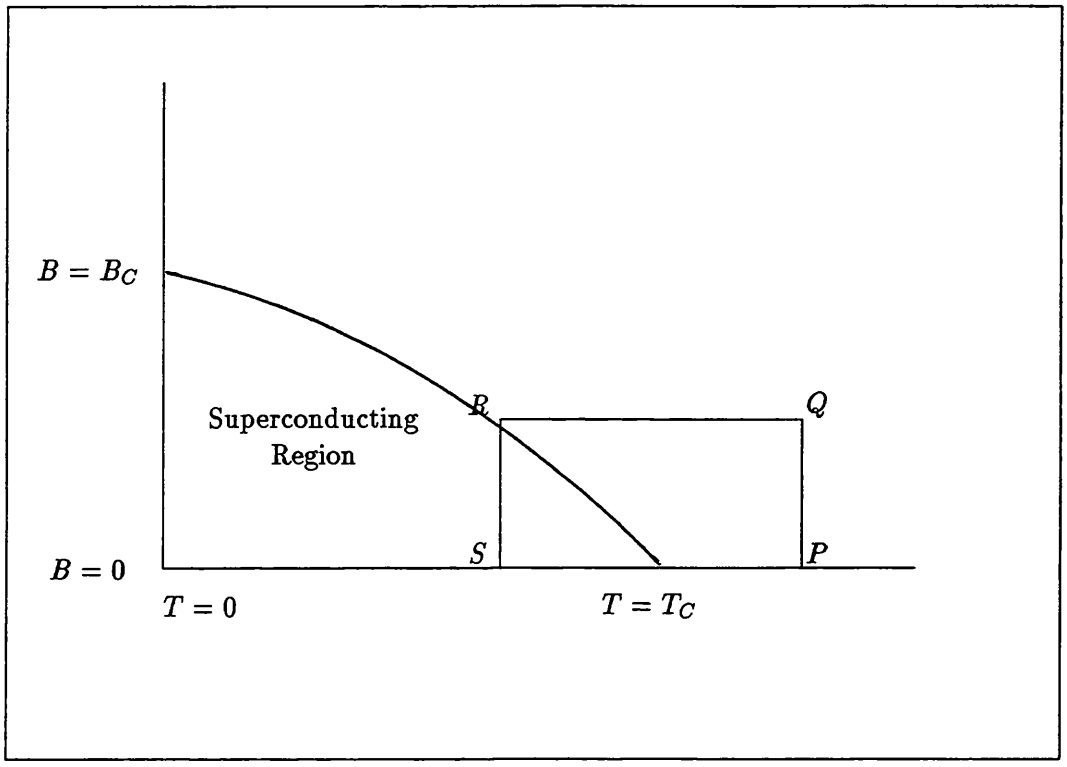


Figure 1.2: The Meissner effect

that if the field is greater than a certain threshold value the superconducting current will virtually come to a halt, as predicted by Onnes. This equation was determined, in the manner described above, by a new model of superconductivity; the model was in its own turn suggested by the phenomena. This reconstruction explains why no satisfactory theory of superconductivity was derived before the discovery of the Meissner effect. A novel conception, embodied in the model of the superconductor as one huge diamagnet, was required for a successful electromagnetic treatment of superconductivity, and such conception was not available before the discovery of the Meissner effect.

1.4.4 The Role of the Theoretical Context

Steven French and James Ladyman[28] have recently made an important contribution to my case study, as part of their response to 'The Tool-Box of Science' [34]. Their work emphasises two aspects of the historical episode. First, French and Ladyman argue that there is a substructure common to the 'acceleration equation' theory and to the London equation. As this substructure is derivable from either theory, it is hence confirmed in both cases. From the acceleration equation (1.4.4), by taking the curl and using the identity $\nabla \times \mathbf{E} = -\frac{1}{c}\frac{d\mathbf{H}}{dt}$, one can derive the following equation:

$$\nabla \Lambda \frac{d\mathbf{j}}{dt} = -\frac{1}{c}\frac{d\mathbf{H}}{dt} \tag{1.4.8}$$

As French and Ladyman point out this equation is the common substructure to both the 'acceleration equation' theory and the Londons' theory. On the one hand (1.4.8) can be derived from the 'acceleration equation' theory as above; on the other hand (1.4.8) also follows from the Londons' theory by simply taking the time derivative of the Londons' equation (equation (1.4.5)). French and Ladyman [68, page 33] conclude that:

> The acceleration equation goes beyond this [(1.4.8)] and in doing so 'expresses a prejudice' in so far as it is not supported

[28] French and Ladyman, 'Superconductivity and Structures: Revisiting the London Account' [68].

by experience. Thus what we have here is an example of what Post[29] calls 'stripping', where the old theory is '...stripped of dispensable, not independently confirmed superstructure'.

As can be seen by inspecting figure 1.2, it follows from equation (1.4.7) that the *dispensable superstructure* is in this case constituted by all those solutions to the 'acceleration equation' that correspond to a non-zero magnetic flux inside the material before the phase transition (i.e. all those cases where $\mathbf{H_0} \neq 0$). Hence, in this case, the 'stripping' of dispensable superstructure would amount to restricting the initial conditions on the old theory (in particular, it would amount to restricting $\mathbf{H_0}$ to $\mathbf{H_0} = 0$). Was the Londons only contribution to restrict the set of initial conditions in the old 'acceleration equation' in order to get rid of the falsified substructures, while retaining the well-confirmed ones? Bardeen[30], for instance, may seem to be saying just this when he writes in his impressive review article:

> The Londons added (1.4.5) to the earlier 'acceleration' theory of Becker, Sauter and Heller to account for the Meissner effect.

But, it is not possible to read Bardeen as claiming that the Londons merely 'stripped away' the dispensable superstructure. The replacement of the set of solutions that involve the initial field in the superconductor by the family of exponential solutions *is not a restriction* of the old theory to the case where there is no initial magnetic flux inside the material, i.e. to the case where the initial external field $\mathbf{B_0}$ is zero. It is true that the 'acceleration equation' theory and the Londons' theory fully agree that in this particular case the magnetic flux inside the superconductor after the transition will vanish. Nevertheless the whole point of the Londons' theory is to show that the flux inside the superconductor is vanishingly small *even if* the initial flux at the time when the transition took place is not zero, i.e. even if the material

[29] Post [111, page 229].

[30] Bardeen [9].

is initially placed in a magnetic field $\mathbf{B_0} \neq 0$. Whenever the magnetic field is not vanishingly small outside the material before the transition the two theories give inconsistent predictions as regards the expulsion of flux: the 'acceleration equation' theory predicts no expulsion, while the new theory predicts a brutal change, consistent with the Meissner effect. The Londons of course accept that in the case $\mathbf{B_0} = 0$ the 'acceleration equation' theory gets it right. But they do not remain silent about those other cases that this theory does not get right. They provide a whole new theory that has the same predictions for the $\mathbf{B_0} = 0$ case, but gives the correct predictions for the other cases. In general, writing down a new equation for the value of a physical quantity in a theory is not equivalent to restricting the initial conditions on the old equations.

The Londons did not just 'strip away' the old 'acceleration equation' theory. They proposed a radically different account that, as I have shown, could not be derived in any way from classical electromagnetism, or for that matter from the 'acceleration equation' itself. French and Ladyman correctly note that the Londons new proposal keeps intact the substructure that generates the predictions that the 'acceleration equation' had got right, while providing new predictions for those cases the 'acceleration equation' could not cope with. This is not much of an argument for or against the idealization account (nor is it intended to be one); rather it brings to the fore the Londons' common sense and their sound scientific methodology.

Secondly, French and Ladyman emphasise the very rich theoretical background to the Londons' theory of superconductivity. In the 'Tool-Box of Science' a distinction was made between *theoretical* and *phenomenological* models; and it was argued that the Londons' account of superconductivity (equation (1.4.5)) constituted a 'phenomenological model'. Indeed the conclusion of the 'Tool-Box' was that:

> ... we have stumbled upon an example of phenomenological model building about which the theory-driven view has little to say. What is needed is the recognition of the independence from theory, in methods and aims, of the scientific activity we have

come to call phenomenological model building. [34, page 11]

French and Ladyman object to this conclusion. They argue that the London account of superconductivity was theoretical, not phenomenological, as its construction was not independent of theory; and they stress that the Londons' work cannot be understood outside its proper theoretical context. As they write:

> The construction of the London-London model did not proceed 'phenomenologically', in the sense of being independent of theory, but rather on the basis of a consideration of the earlier theoretical context and, in particular, of what could be retained from that context *in the light of Meissner's work.* (op. cit. [68, page 34], my italics)

According to French and Ladyman, there are two salient aspects of this background theoretical context. First, of course, there is the 'acceleration equation theory' of Becker, Sauter and Heller, which was "*the first attempt to supplement Maxwell's equations*" (Gavroglu, [71, page 118]). Second, French and Ladyman claim, on the basis of Dahl's[31] recent exposition, that the model of a superconductor as a diamagnet was already 'in the air', and it was not introduced for the first time by the Londons in their 1934 paper. Fritz London himself, in a passage in a later paper[32] in 1935, acknowleged this when he wrote:

> It is rather seductive to consider the supra-current as a kind of diamagnetic current, an idea which has sometimes been uttered in the past, *now more seductive than ever, since Meissner's experiment seems to reveal to us the more elementary phe-*

[31] Dahl [40].

[32] I want to thank Steven French for calling this passage to my attention, and for sending me with a copy of Fritz London's paper which is absurdly unavailable in Oxford's Bodlean library.

nomenon to which one may hope to reduce the so enigmatical phenomenon of conductivity. (op. cit. [97, page 26], my italics)

Indeed a couple of years earlier, in a letter to *Nature*, Gorter[33] had suggested that if diamagnetism was regarded as the general defining characteristic of superconductors then the phase transition would be reversible; thus opening the way for a consistent thermodynamics of superconductivity. The Londons took up the suggestion and provided an electromagnetic treatment of superconductivity by correcting the 'acceleration equation' theory in light of the diamagnetic model. As French and Ladyman note:

> It was Gorter who realised, *on reading of Meissner's results*, that the contradiction between the thermodynamic and magnetic representations of the superconducting transition could be dissolved if the condition of $\mathbf{B} = 0$ (...) were regarded as the general defining characteristic of superconductors. The task was then to reconfigure the electromagnetic treatment in line with this criterion for distinguishing the superconducting phase, which is precisely what London and London did. (op. cit. [68, page 27], my italics)

If a 'phenomenological model' is a representation, or a collection, of empirical data that involves no theoretical assumptions or concepts, the Londons account of superconductivity was not a phenomenological model. It was a heavily theory-laden account. It was in fact an application of classical electromagnetism. The distinction between 'phenomenological' and 'theoretical' models in no way ought to presuppose a notion of non theory-laden observation. Nor is 'phenomenological model' intended as a model of raw uninterpreted empirical data. Instead the distinction between phenomenological and theoretical models outlined in the 'Tool-Box of Science' was intended to turn solely upon the 'theory-driven' view of model-building. A theoretical model is accordingly defined as a model that can be derived

[33] Gorter [77].

from another theoretical model by the introduction of corrections suggested by the theory. As we saw in the discussion of the idealization account, it is essential that the criterion whereby we choose corrections figures already in the theory; otherwise the resulting model would be called 'phenomenological'.

The terminology of 'theoretical' and 'phenomenological' models was certainly not used by the Londons themselves, nor was it known to them; and it led French and Ladyman to infer, incorrectly, that the claim of 'Tool-Box of Science' was that the Londons' theory was inductively found, a claim which would of course be false. As French and Ladyman write:

> Obviously model construction, and theory change in general, does not proceed by logical derivation –how would there be any genuine change if it did?– nor does it proceed purely phenomenologically, which is to say inductively (op. cit. [68, page 34]).

It should be clear that I am not arguing that the Londons found their theory by inductive means. On the contrary, they found it by introducing corrections into the 'acceleration equation' theory. The point is that these corrections were prompted by an independently standing model of the superconductor as a diamagnet, –a model that was not itself suggested by theory. Hence, the London equation was not a 'theoretical' description, in the sense that I have been using the term here: the derivation of the London equation was not 'theory-driven'[34].

My thesis is then that the Meissner effect acted as the *catalyst* for the diamagnetic model of superconductivity. Of course the diamagnetic model also served to solve other long standing problems, mainly the inconsistency of the

[34]See the discussion of McMullin's views on application in section 1.2.2. Incidentally note that French and Ladyman's remark that "*obviously model construction [...] does not proceed by logical derivation*", if indeed addressed to the theoretical/phenomenological models distinction, misses the target as a critique of McMullin's views on application. For 'theory-driven' is not as strong as 'entailed by theory'. If the corrections are logically dictated by the theory then, of course, they are also *suggested* by it. But a theory can suggest corrections that it does not logically entail.

electromagnetic and the (reversible) thermodynamic treatments of superconductivity. But this inconsistency could not have, on its own, suggested the diamagnetic corrections to the 'acceleration equation'. Historically the discovery of the Meissner effect signalled the turning point. Before Meissner's discovery the diamagnetic model was hardly credible; the analogy with ferromagnetism held tremendous sway, as it was so much more naturally fitted to perfect conductivity; but after Meissner's discovery the diamagnetic model gained an irresistible appeal. Whether it was the Londons or Gorter who initially suggested the model is of secondary importance here. It would not even matter if the model had been mentioned before Meissner's discovery. What matters is that for the community working in the field the transition from the ferromagnetic to the diamagnetic conception of superconductivity was prompted by the discovery of the Meissner effect. In fact the quoted passages (from French and Ladyman, Fritz London, and French and Ladyman again) make my point most forcefully. Every one of the emphasised phrases in those passages unambiguously states the fact that the diamagnetic model of superconductivity was prompted by Meissner's discovery.

Kostas Gavroglu's recent biography of Fritz London provides further evidence, once again in the form of testimony from Fritz London himself. In a 1937 letter to Casimir, London –while acknowledging the importance of the theoretical context– makes it clear that the Meissner effect was the determining factor in the acceptance of the diamagnetic conception:

> The paper by Gorter, and that of Gorter and yourself, made a strong impression on me at the time, and incited me to engage myself with superconductivity. It is true that the Meissner effect could have been predicted from Gorter's ideas. The fact that the development of things did not take place this way and needed an experimental push, always appeared to me a sign that the acceptance of the *reversibility* was not at all self-evident, because of the fact that all the experiments displayed hysteresis and other non-reversibilities. It was at that time only a *hypothesis* that was constructed in the dark ... and it was not proper to

> interpret in *exactly* that manner the objective non-reversibility and to pinpoint in *that* particular manner the assumption of reversibility. This is why the verification of this magnetothermic phenomenon seems to be so important for me. Because it means something more than a mere verification of thermodynamics. It also means, as far as I know, the verification that what we *assumed* to be reversible is *indeed* reversible. ([71, page 121], italics in the original).

To sum up, French and Ladyman's emphasis on the important role played by the theoretical context in this episode is in general agreement with my reconstruction, and their work in fact supports my main claim. I claim that the Londons' account, even if not a deidealization, constituted an application of electromagnetic theory to superconductivity. For we arrive at the London equation (equation 1.4.5) by introducing corrections on a previous electromagnetic description of the phenomenon: the 'acceleration equation'. So I welcome a study of this episode that emphasises the Londons' background in electromagnetic theory, and in particular, their debt to the 'acceleration equation' theory of Becker, Sauter and Heller. I go further than French and Ladyman in claiming that, due to the fact that the diamagnetic conception was ultimately impressed upon the community by experimental facts independent of theoretical considerations, the Londons' account is no 'deidealization' of electromagnetic theory. The theory did not suggest, and could not have determined, how to correct the 'acceleration equation' to account for the Meissner effect. I have no reason to believe that French and Ladyman would disagree with this claim. On the contrary, as the emphasised passages show, their work not only fails to undermine my claim, it rather explicitly supports it.

1.5 Application in Practice: Problems for Realism

In providing a macroscopic description of the Meissner effect in electromagnetic terms, the Londons effectively succeeded in providing a satisfactory

application of electromagnetic theory to superconductivity. However, they did not *deidealize* electromagnetic theory. Instead they came up with a model that permitted them to impose a novel constraint upon the original theoretical construction. This case study is not exceptional; on the contrary, *many* scientific applications are derived in this way. In astrophysics, for example, there are several models of stellar structure. A certain conception of the internal constitution of a star, which determines the form of the convection forces in the stellar plasma, has to be assumed before the quantum theory of radiation can be applied. For each different conception there is a corresponding application of the theory, a family of models, that could not have been derived from the theory alone. The idealization account is then not a universal account of scientific theory-application. It is far too restrictive. It imposes constraints so severe that they are not always –indeed are rarely– met in practice.

1.5.1 The Epistemology of Theory-Application

What are the epistemological implications of the rejection of the idealization account? I shall focus the discussion closely upon the case study. The Londons built an application of electromagnetic theory to superconductivity; and yet, on McMullin's account, the theory was in no way confirmed by the phenomenon of superconductivity. Confirmation requires that the theory itself must suggest the introduction of corrections into the theoretical description. For, as McMullin points out[35], a theoretical description is *ad hoc* with respect to a theory that does not suggest or motivate its derivation; and an *ad hoc* description, or hypothesis, cannot increase the degree of confirmation of a theory with respect to which it is *ad hoc*[36].

[35]See the discussion in section 1.2.2 in this Thesis, and in particular the passage quoted from McMullin.

[36]Hempel too makes this claim in his [84, pp.28-30], although he there ascribes a slightly different meaning to the term *ad hoc*. For Hempel, a hypothesis is *ad hoc*, with respect to some theory, if it has no surplus empirical content over the theory other than the particular phenomenon that it is specifically called to account for.

In constructing their account of superconductivity, the Londons introduced a correction into the previously available theoretical description. The correction was certainly not arbitrary, since it was justified by a new model of superconductivity. However, this model was not suggested by the theory –it was suggested by a newly discovered physical effect. On McMullin's confirmation theory, classical electromagnetism was not in this instance genuinely confirmed at all. Was it *neither* confirmed *nor* disconfirmed, or was it simply disconfirmed? The answer depends on what we take electromagnetic theory to be *circa* 1933.

There are two possible pictures. It is possible to take 'electromagnetic theory' in an extended, historical sense, as constituted by all applications to electromagnetic phenomena known to the Londons. The 'acceleration equation' is part of electromagnetic theory, when construed in this extended sense. And, as we saw, the Londons gave an account that contradicted the acceleration equation predictions in a range of cases. Hence, if taken in this 'historical' sense, classical electromagnetism was indeed *disconfirmed* by the Meissner effect.

Alternatively, one may provide an abstract reconstruction of electromagnetic theory. The standard reconstructions normally assume that classical electromagnetism is constituted by the deductive closure of Maxwell's equations. Now, the 'acceleration equation', although not inconsistent with Maxwell's equations, is not a logical consequence of these equations. It can be postulated alongside them, in just the way Ohm's law is often postulated alongside Maxwell's equations, but it cannot be derived from them. Nor is the Londons' account a logical consequence of Maxwell's equations; although it is also consistent with them, and can be postulated alongside them[37]. Thus, neither the 'acceleration equation' nor the London's equation is part of electromagnetic theory, understood in this abstract manner. And it follows that, in this abstract reconstruction, the Londons account

[37]N.B. It is perfectly possible for a theory T to be consistent with each of two mutually inconsistent assumptions a and b, –as long as T entails neither a nor b, of course.

provided neither a confirmatory nor a disconfirmatory boost for classical electromagnetism.

And yet, the Londons' treatment did increase scientists' confidence in electromagnetic theory. Superconductivity had proved difficult to model in classical electromagnetism for a long time, and many were beginning to despair that a consistent electromagnetic treatment would ever be found. The diamagnetic conception played a key role in the Londons' explanation of the phenomenon of superconductivity, which reveals the extent to which a mediating model carries genuine physical knowledge. The Londons' theory was generally accepted to account rather accurately for the rate of magnetic flux expulsion from a superconductor during the phase transition reported by Meissner and Ochsenfeld in their experimental investigations[38]. From this application of electromagnetism we learn that superconductivity is an essentially diamagnetic effect; that a superconductor is not a ferromagnet; and, moreover, as the Londons' account correctly predicts the rates of expulsion of magnetic flux observed by Meissner and Ochsenfeld, we gain a theoretical understanding of the Meissner effect. The Meissner effect does not appear as a mysterious side-effect of superconductors; instead it takes centre stage, it becomes a fundamental hallmark of superconductivity.

The Londons' account of superconductivity provided an extra 'boost' of confidence in classical electromagnetism which the old 'acceleration' theory could not provide. But, as we have seen, on McMullin's idealization account of application, the Meissner effect does not make electromagnetic theory more likely to be true. It seems that this extra boost of confidence in electromagnetism cannot be captured by the standard realist theory of confirmation, so I shall refer to the kind of support that the Londons treatment provided for electromagnetism as *degree of confidence* rather than degree of confirmation.

The fact that the Londons' equation accounts for the Meissner effect

[38]Although there was some initial resistance to the London's theory on empirical grounds. In particular Von Laue disagreed; for the dispute between Fritz London and Von Laue, see Gavroglu [71, pages 123-127].

gives grounds to believe that classical electromagnetism is *instrumentally reliable*. But it does not constitute evidence for the truth of classical electromagnetic theory. Here *degree of confidence* and *degree of confirmation* seem to depart. Degree of confidence, unlike degree of confirmation, does not point to the likelihood of the theory to be true; it only to points to the reliability of the theory as an instrument in application. The theory is a reliable instrument if it is capable, perhaps when conjoined with good enlightening mediating models, of generating successful applications. And from the fact that the theory is instrumentally successful, the truth of the theory does not follow.

Or does it? Would it not be a miracle if the theory was false, yet instrumentally successful? Does the instrumental success of scientific theories not argue for scientific realism? Arguments of this kind in favour of realism are, of course, well known in the literature[39]. Typical antirealist responses to this argument are equally well known. For instance Arthur Fine[40] responds that the 'no-miracles' argument is riddled with circularity: it assumes that precisely the very sort of inference from explanatory power to truth that realism sanctions and instrumentalism contests for scientific practice, is valid at the 'meta-level' and can be used as part of an argument for realism in general. As a response, scientific realists have turned to the pragmatic virtues of realism, and they have tried to show that no version of antirealism is in any better shape. In particular the debate has focused upon Bas Van Fraassen's version of antirealism, known as *constructive empiricism*[41].

It is not necessary to rehearse these arguments here. The issues about realism that I am raising are tangential to the recent debate between sci-

[39]The original 'no-miracle' arguments are due to Putnam [113], and Boyd [16], and [17].

[40]Fine, 'Unnatural Attitudes: Realist and Instrumentalist Attachments to Science' [63].

[41]For Van Fraassen's constructive empiricism see 'To Save the Phenomena' [132], reprinted with corrections in his book *The Scientific Image* [133]. A collection of papers by critics of constructive empiricism, together with responses by Van Fraassen is contained in *Images of Science* [37].

entific realists and constructive empiricists. The contrast between *degree of confidence* and *degree of confirmation* is not captured by the debate. Degree of confirmation measures the degree of a theory's empirical adequacy, or its truth. Degree of confidence, as I would like to define it, is not grounded on an evidential relationship of the truth-conferring type between a theory and phenomena. Increased *confidence* in classical electromagnetism need not to be accompanied by an increase in one's estimated probability that it correctly describes the world, i.e. that it is true. The success of the London model does not provide warrant for that. Neither does it warrant an increase in one's estimated probability that the theory correctly describes the phenomenal world, ie. that the theory is empirically adequate. Unlike degree of confirmation, *degree of confidence* is not a function of a theory's predictive power and empirical adequacy. It is not evidential but pragmatic, a function of the success of a theory in generating applications to diverse phenomena whenever conjoined with the appropriate mediating models. I call this feature of theories 'instrumental reliability' in order to distinguish it sharply from empirical adequacy. The instrumental reliability of a theory does not require, nor does it necessarily follow from, its empirical adequacy. This is of course in agreement with my argument so far: the fact that classical electromagnetic theory can be applied to superconductivity should not be taken as an indication that the theory is *true* to superconductivity phenomena.

1.5.2 Conclusions

Let us grant that scientists do see a theory's applications as providing some degree of confidence in the theory. Does this not argue for the idealization account, and hence for the realist epistemology that underpins it? Some scientific realists, such as McMullin, think so. The idealization account, they think, is *required* by scientific epistemology. On the idealization account the explanatory power of a theory is exhibited through its applications, and the theory is more likely to be true in view of the success of its applications. So, realism is required to make sense of the epistemology of theory-application.

However, the instrumentalist can give a similarly good account of the

epistemology of application. Scientists' increased confidence in a theory that has generated many applications is a result of the theory's instrumental success in modelling the phenomena. This gives, at most, confidence that the theory will continue to generate successful applications in the future, i.e. that it is an *instrumentally reliable* theory. And this argues against realism: a successful application of a theory need not constitute evidence that the theory is true. The applications of a scientific theory do not necessarily yield the kind of evidential support for the truth of the theory that scientific realism requires them to.

1.6 Summary

In this chapter I have introduced the notion of a mediating model (section 1). By means of a case study in superconductivity I have argued that mediating models play a key role in the application of scientific theories (section 4). I have described McMulllin's realist account of scientific theory application (section 2), and I have argued that the role of models as mediators is incompatible with such an account (section 3). I have finally argued (section 5) that the pervasive role of models as mediators strongly suggests a very different account of application, one that carries no commitment to a realist epistemology. Yet this alternative account can also explain why, in scientific practice, confidence in a scientific theory is typically boosted by its applications.

Chapter 2

The Semantic View: Empirical Adequacy, Truth and Application

2.1 To Save the Phenomena

In *The Aim and Structure of Physical Theory* [49], Pierre Duhem argued that the aim of physics is 'to save the phenomena'. The physicist's task is to construct physical theories that account for the phenomena, in the following two ways. Firstly, theories provide scientists with a Machian economy of thought that permits them to hold in mind a number of empirical regularities all at once:

> Instead of a great number of laws offering themselves as independent of one another, each having to be learnt and remembered on its own account, physical theory substitutes a very small number of propositions, viz., fundamental hypotheses. [...] Such condensing of a multitude of laws into a smaller number of principles afford enormous relief to the human mind, which might not be able without such an artifice to store up the new wealth it acquires daily. (Duhem, *Aim and Structure* [49, page

21])

Secondly, theories contain only the most abstract principles which can classify and impose structure upon the diversity of natural phenomena:

> Experimental physics supplies us with laws all lumped together and, so to speak, on the same plane, without partitioning them into groups of laws united by a kind of family tie. [...] On the other hand theory, by developing the numerous ramifications of the deductive reasoning which connects principles to experimental laws, establishes an order and a classification among these laws. [...] Theory gives, so to speak, the table of contents and the chapter headings under which the science to be studied will be methodologically divided, and it indicates the laws which are to be arranged under each of these chapters. (Duhem, op.cit. [50, pp. 23-24]).

The physical laws that theoretical hypotheses classify are not themselves empirical regularities, but rather inductive generalisations of empirical regularities. Observed regularities necessarily have only a finite number of instances, –as they are constituted by a collection of concrete facts, normally relating to past spatio-temporal coincidences between particular kinds of events–, while laws have a potentially infinite number of instances. Hence the observed regularities do not fix the physical laws. In modern philosophical jargon: physical laws are underdetermined by the phenomena. Similarly, for Duhem, physical theory is underdetermined by the set of physical laws. As a matter of principle there will be several hypotheses that can equally well classify the set of physical laws, and equally well account for the phenomena.

According to Duhem, the truth of a theory can only manifest itself in the theory's capacity to account for the phenomena:"*agreement with experiment is the sole criterion of truth for a physical theory*" (Duhem, op.cit. [49, page 21]). However, two empirically equivalent theories may postulate radically different properties of the entities and processes that underlie the

phenomena. (The Ptolemaic system of the world, for instance, postulates that the earth is static at the centre of the Universe, while in the Copernican system the earth follows a perfect circular motion around the sun.) This yields the well-known sceptical argument from underdetermination: How are we to choose, among all these empirically equivalent possibilities, the one and only true theory? If the truth of the theory manifests itself only in its capacity to save the phenomena, it is not possible to select the true theory from among the set of empirically indistinguishible hypotheses. The assertion that one of them is true becomes an empty metaphysical claim, devoid of empirical content[1].

The only *necessary* requirement for accepting a theory is that it must save the phenomena. But how exactly is a theory supposed to 'save the phenomena'? How can physicists know when the theory has achieved its aim to 'save the phenomena'? And how much of the phenomena is a theory supposed to 'save'? In a collection of papers originally published in 1908, and now available under the title *To Save the Phenomena* [50] Duhem suggests that astronomy, in the tradition of Eudoxus and Ptolemy, will provide the model: theories save the phenomena in just the same fashion astronomical hypotheses describe the observed motions of the objects in the heavens[2].

[1]This doctrine is essentially of scholastic origin; Duhem (op.cit. [50, page 41]) approvingly quotes Saint Thomas Aquinas:

> Astronomers have tried in diverse ways to explain this motion [of the planets]. But it is not necessary that the hypotheses they have imagined be true, for it may be that the appearances the stars present might be due to some other mode of motion yet unknown by men.

[2]In *To Save the Phenomena* [50] Duhem describes two competing astronomical traditions. In the tradition of Eudoxus and Ptolemy celestial phenomena are 'saved' if the motions of objects in the heavens can be calculated, and predicted. In the tradition of Aristotle and Posidonius some further conditions must be satisfied: hypotheses about the motions of the objects in the heavens must be based upon the solid principles of physics. The Aristotelian tradition is lost for centuries to the Christian world, preserved only in the Arab writings of Averroes and Al-Bitrogi, translated into latin only late in the middle ages in the Court of Alfonso X of Castille; it is adopted during the Renaissance by the Italian Averroists of the School of Padua and, according to Duhem, it is inherited by Coperni-

The observable predictions of the theory are found by deduction from first premises expressing nomological relations between physical quantities, together with boundary conditions and a number of auxiliary assumptions about the workings of instruments, etc. These predictions must be borne out if the theory is to 'save the phenomena'. This hypothetico-deductive methodology is of course not in contradiction with Duhem's belief in underdetermination. An experimental contradiction of a theoretical prediction does not necessarily result in a refutation of the theory, as the underdetermination argument still applies in its holistic form:

> The physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (Duhem, op.cit. [49, page 187]).

What is interesting about Duhem's later historical work is that in the astronomical model both the application and the testing of astronomical hypotheses follow the hypothetico-deductive method. In order to test a hypothesis about the constitution of the heavens, we look for a derivation from the hypothesis of a sequence of positions of a planet, given the appropriate boundary conditions, and we test that sequence by direct observation. Equally a model for the motion of a planet is a sequence of positions deduced from a hypothesis, in just the same manner. But, in general, the distinction between confirmation and application is coherent, and it is important.

---

cans and Inquisitors alike (although with obviously competing interpretations) during the Galilean trials. The former tradition, which Duhem favours, begins with Eudoxus and Ptolemy; it is pursued in Christianity by the scholastics of the University of Paris during the thirteenth and fourteenth centuries and their followers in German-speaking countries thereafter; and it eventually gives rise to the conciliatory and ecumenical views of Osiander and Bellarmino.

In fact, Duhem himself had carefully distinguished between what he called experiments of *testing* and experiments of *application*:

> You are confronted with a problem in physics to be solved practically; in order to produce a certain effect you wish to make use of knowledge acquired by physicists; you wish to light an incandescent bulb; accepted theories indicate to you the means you have to secure certain information; you ought, I suppose, to determine the electromotive force of the battery of generators at your disposal; you measure this electromotive force: that is what I call an experiment of application. This experiment does not aim at discovering whether accepted theories are accurate or not; it merely intends to draw on these theories. In order to carry it out, you make use of instruments that these same theories legitimise; there is nothing to shock logic in this procedure. (Duhem, op.cit. [49, page 184]).

In *To Save the Phenomena* Duhem assimilated the notion of application to the notion of empirical adequacy: in the model provided by astronomy, the domain of application of a theory (such as, for instance, Ptolemy's) coincides with its domain of empirical adequacy. (If Ptolemy's theory can be applied to all kinds of observed motions in the sky, then it is an empirically adequate theory of celestial motion.) In other words, the cases that confirm the theory are precisely those to which the theory gets applied. This view, which construes a theory's applications as strict deductions from theory, is not, however, a consequence of Duhem's empiricist epistemology. It is instead a consequence of taking astronomy to provide the model for the application of theories. Astronomy provides a handy picture, one in which application and confirmation go hand in hand. But there is no reason why physics as a whole ought to accord to this picture, and there is no reason why empiricism ought to be committed to it.

2.2 The Nature of Scientific Theories

> The syntactic picture of a theory identifies it with a body of theorems, stated in one particular language chosen for the expression of that theory. This should be contrasted with the alternative of presenting a theory in the first instance by identifying a class of structures as its models. In this second semantic, approach the language used to express the theory is neither basic not unique; the same class of structures could well be described in different ways, each with its own limitations. The models occupy centre stage.
>
> (Van Fraassen, *The Scientific Image* [133, page 44])

2.2.1 The Syntactic Conception

With a few remarkable exceptions the logical positivist tradition focused on the empirical adequacy of scientific theories, and construed empirical adequacy as a logical relation between a theory and the phenomena that the theory describes. In Carl G. Hempel's early work the empirical adequacy of a theory is closely connected with the meaning of its theoretical terms. Theoretical terms are supposed to gain their meaning from the observational vocabulary to which they are linked by *correspondence rules* –sentences that contain a mixture of theoretical and observational terms. In 'The Theoretician's Dilemma' Hempel writes:

> A deductive system can function as a theory in empirical science only if it has been given an *interpretation* by reference to empirical phenomena. We may think of such interpretation as being effected by the specification of a set of *interpretative sentences*, which connect certain terms of the theoretical vocabulary with observational terms. (Hempel, *Aspects of Scientific Explanation* [83, 184])

According to the Received View of scientific theories, a theory is a set of sentences in a first-order language L. The logical vocabulary consists

of the usual first-order logical constants, while the non-logical vocabulary is divided into V_O which contains the 'observational' terms, and V_T which contains the 'theoretical' or non-observational terms. This splits the language L into three sublanguages: L_T, the theoretical language, which contains sentences formed exclusively out of terms in V_T; L_O, the observational language, whose sentences are formed out of terms in V_O only; and L_M, the mixed language, which has 'mixed' sentences containing at least one term of V_O and at least one of V_T. The observational language is then given a full semantic interpretation by describing a domain of purely observable entities and relations, and by setting up a mapping from the terms in V_O into a set of concrete observable events or entities, and from the predicates that appear in the sentences of L_O into directly observable relations between those events or entities. It then becomes possible to give a *partial interpretation* of the theoretical language L_T by providing two kinds of postulates: theoretical postulates and mixed postulates.

Theoretical postulates are sentences in L_T which implicitly define some of the terms in the theoretical vocabulary by means of antecedently understood terms in V_T. The set of theoretical postulates T is said to constitute the 'axioms' of the theory. But if these were the only sentences in the theory, all the terms in V_T would be implicitly defined by the postulates in T, which would make T, the theory, analytically true and, therefore, true a priori[3]. This consequence may, *pace* Quine, be welcomed in the context of pure mathematics, but it is unacceptable for scientific theories. Scientific theories therefore necessarily contain further sentences that connect the observational and theoretical vocabularies –for no scientific theory can the 'mixed' language L_M be empty.

The sentences in the mixed language L_M are called mixed postulates, or *correspondence rules.* Let us refer to the set of all correspondence rules as C; a scientific theory is then defined, in the syntactic conception, as the union of the set of theoretical postulates and the set of correspondence

[3]Essentially this point is made by Hempel in [86, page 248].

rules, and can be denoted by TC. Correspondence rules have a semantic, cognitive and empirical role. They provide a partial semantics for theoretical terms; they guarantee that the propositions asserted by the theory are truth-apt and receive a truth value (which one depends upon the truth value of the observational sentences under their ascribed interpretation); and they provide the empirical import of the theory. As Frederick Suppe writes:

> Correspondence rules serve three functions in the Received View: first, they define theoretical terms; second, they guarantee the cognitive significance of theoretical terms; third, they specify the admissible experimental procedures for applying theory to phenomena. (Suppe, *The Structure of Scientific Theories* [124, page 17])

A theory is empirically adequate if what it entails about the phenomena is indeed the case. What does the theory entail? Well, according to the syntactic conception of scientific theories, this has an easy answer. We began precisely by assuming that the observable content of the theory can be isolated in a sublanguage of the language of the theory, L_O, which receives a full semantic interpretation in terms of observable processes and entities. Hence the empirical content of the theory is the set of all synthetic sentences in L_O. If these are all true, the theory is empirically adequate. This is, in the syntactic conception, the basic criterion for the empirical adequacy of a scientific theory.

2.2.2 Critique of the Syntactic Conception

It is not necessary to review here the history of misery and pitfalls that followed the introduction of the distinction between the theoretical and the observational vocabulary of a theory. The many and varied criticisms of the distinction –due mostly to Norwood Russell Hanson, Paul Feyerabend, Hilary Putnam, Grover Maxwell[4] –turned on aspects of the theory-ladenness

[4] An overview and summary is in Suppe, op.cit. [124].

of observation. Suffice it to say that Hempel himself famously abandoned the requirement that the division between subvocabularies be drawn on any such distinction. Instead he adopted a division between 'new' and 'antecedently understood' terms, and replaced correspondence rules with *bridge principles*:

> The bridge principles will evidently contain both the terms of [the new vocabulary] and those of the vocabulary used in formulating the original descriptions of, and generalisations about, the phenomena for which the theory is to account. This vocabulary will thus be available and understood prior to the introduction of the theory, and its use will be governed by principles which, at least initially, are independent of the theory. Let us refer to it as the *pretheoretical* or *antecedent vocabulary*, relative to the theory in question. (Hempel [85, page 143])

This vocabulary is *pretheoretical* relative to a theory, in the sense that its terms already appear in previous theories. Terms in the antecedent vocabulary need not be linked to observational terms, nor are they to receive a full semantic interpretation by means of observable entities and processes:

> The terms of the antecedent vocabulary are by no means assumed to be "observational" in the sense of the familiar theoretical - observational distinction; that is, they are not required to stand for entities or characteristics whose presence can be ascertained by direct observation, unaided by instruments and theoretical inference. (Hempel [86, page 245].)

Hempel's revision was designed to patch up the syntactic conception *as a theory of meaning* for theoretical terms. In abandoning correspondence rules and adopting bridge principles instead, Hempel effectively abandoned the symmetry between theoretical meaning and empirical adequacy. Bridge principles connect the new terms with antecedently understood terms, providing a partial interpretation of the new terms. But as the antecedent vocabulary is not restricted to observable entities and processes, bridge

principles cannot give empirical content to the theory; the restriction of the theoretical postulates to the antecedent vocabulary does not yield the empirical import of the theory. Therefore, Hempel's revision leaves us in the dark as to how to characterise the empirical adequacy of a theory. In fact, this revision reflects Hempel's conviction that such a characterisation is not needed, after all, for a satisfactory understanding of the meaning of theoretical terms.

Bas Van Fraassen has provided independent arguments against the syntactic characterisation of empirical adequacy. In his book *The Scientific Image* he argues that it is impossible to isolate the empirical import of a theory in a purely syntactic fashion, by drawing a distinction between theorems in terms of vocabulary. The empirical content of a theory is what the theory says about what is observable, and nothing more. However, many theoretical commitments of a theory can be expressed in its restricted 'observational' vocabulary:

> Any unobservable entity will differ from the observable ones in the way it systematically lacks observable characteristics. As long as we do not abjure negation, therefore, we shall be able to state in the observational vocabulary (however conceived) that there are unobservable entities, and, to some extent, what they are like. The quantum theory, Copenhagen version, implies that there are things which sometimes have a position in space, and sometimes have not. This consequence I have just stated without using a single theoretical term. (Van Fraassen [133, page 54])

The syntactic characterisation of empirical import misfires. In the syntactic conception the empirical import of a theory is constituted by the theorems of the theory expressed in the observational vocabulary: TC/V_O. Two theories TC_1 and TC_2 are considered to be empirically equivalent if their restrictions to the observable vocabulary are identical, i.e. if $TC_1/V_O = TC_2/V_O$ (assuming that the extension of V_O is the same for both theories). However, the requirement that all sentences in the observational language be true for the

theory to be empirically adequate is too strong. And the criterion of empirical equivalence is correspondingly too strong: theories that are in practice empirically indistinguishible will, according to the syntactic criterion, be judged empirically non-equivalent. Van Fraassen gives several examples of empirically indistinguishible theories that turn out, on the syntactic criterion, to be 'empirically non-equivalent'. For instance, Newton's theory of mechanics and gravitation, with the postulate of absolute space, is indistinguishible from a version of the Leibnizian theory that simply asserts that Newton's theory is empirically adequate. On the syntactic account, however, these theories are not empirically equivalent, because it is possible to make assertions about the properties that absolute space *lacks* within the observational vocabulary of Newton's theory.

The problem is that the restriction of the theorems of the theory to the 'observable' vocabulary does not need to correspond to a demarcation of those claims that the theory makes about the observable part of the world. As Van Fraassen notes:

> The reduced theory $[TC/V_O]$ is not a description of part of the world described by TC; rather $[TC/V_O]$ is, in a hobbled and hamstrung fashion, the description by TC of everything. (Van Fraassen, op.cit. [133, page 55])

Some of the claims that a theory makes about the non-observable part of the world can be expressed wholly in the observable vocabulary. The theory-ladenness of observation points to the fact that the the converse must also be the case: in order to assert the observational consequences of the theory we will often need to appeal to the theoretical vocabulary. So, the division between 'observable' and 'theoretical' vocabularies of a theory cannot track the distinction between those claims that the theory makes about the observable part of the world, and those it makes about the unobservable part. The syntactic conception has some other severe problems (for instance: it makes theories essentially language-relative objects), but I do not need to discuss them here. Like Van Fraassen, I look for an account of theories that can explicate the notion of empirical adequacy. It seems it is impossible to

isolate the empirical content of a theory in a syntactic fashion, by means of a division of the vocabularies in which the theory's claims are cast. So, an alternative, non syntactic, account of scientific theories must be sought. I turn now to Van Fraassen's favourite account: the semantic conception.

2.2.3 The Semantic Conception

The semantic view of scientific theories has antecedents in work by Beth and Von Neumann in the 1930s, 1940s, and 1950s. It was developed into a comprehensive account of scientific theories by Patrick Suppes in a series of papers in the late 1950s[5]. Van Fraassen expresses the basic tenet of the semantic view as follows:

> *To present a theory, we define the class of its models directly,* without paying any attention to questions of axiomatisability, in any special language, however relevant or simple or logically interesting that might be. And if the theory as such, is to be identified with anything at all – if theories are to be reified– then a theory should be identified with its class of models. (Van Fraassen, *Laws and Symmetry* [137, page 222])

There are, however, many kinds of models. What sort of models is the semantic conception appealing to? Defenders of the semantic conception seemingly differ on their answers to this question. I will begin by distinguishing two kinds of models. I call the first *semantic models*, for lack of a better term. These models are also sometimes referred to as *interpretations*, although this terminology is somewhat misleading: a semantic model is not only an interpretation, but an interpretation *and* a structure. The structure is composed of a domain D of objects and some relations R_i defined over the elements in that domain, and can be denoted as $S =< D, R_i >$. A language L is specified which contains logical constants, names and n-place

[5]Suppes [125], [126], and also a manuscript circulated under the title 'Set-theoretical Structures in Science' [128].

predicates. An interpretation of the language L in terms of the structure S is a mapping from the constants in a language into the elements of the domain D, and from the predicates in the language into the relations R_i in the structure. In order to identify a particular interpretational mapping we need to specify both the structure S that constitutes its range, *and also* the language L, including the syntax, that constitutes its domain. Thus, the definition of a semantic model is always tied to a particular language.

The second kind of models I call simply *structures.* To specify these we need not specify a particular syntax; structures are not necessarily tied up to any language. These are typically mathematical structures, defined over a domain of mathematical objects. In Van Fraassen's version of the semantic conception a scientific theory is a collection of models of this kind:

> The impact of Suppes's innovation is lost if models are defined, as in many standard logic texts, to be partially linguistic entities, each yoked to a particular syntax [...] Models are mathematical structures, called models of a given theory only by virtue of belonging to the class defined to be models of that theory. (Van Fraassen, op.cit. [137, page 366])

This suggests that the name of the 'semantic' conception of scientific theories is perhaps misleading. For this name brings to mind the syntax/semantics distinction in linguistics, and the notion of an interpretation of a particular language; while, as a matter of fact, the 'semantic' conception defends the view that scientific theories are families of mathematical structures – and these are not to be thought of as providing an interpretation of any language. Indeed, for very much these reasons, Ronald Giere has recently proposed changing the name of the semantic view to the *model-theoretic* view[6]. The 'model-theoretic' or, as I will call it, the 'structuralist' construal of the semantic conception has an important advantage over the 'semantic-model' construal. It provides an answer to the charge, by Michael Friedman

[6]Giere [74, page 277].

[69] and John Worrall [144], that there is no essential superiority of the semantic over the syntactic conception. If the models the semantic view refers to were models of the first kind, i.e. *semantic models*, the semantic and syntactic characterisations of a theory would be provingly equivalent. Given an elementary class of models (call it N) we know, by the completeness theorem of first-order logic, that there must be a set of axioms Ax, in the language of first order logic, that is satisfied by this and only this set of models. So talk about the models (*semantic* models) is equivalent to talk about the axioms (in the language of first order logic). As Michael Friedman puts it:

> Let us follow van Fraassen in identifying a theory with a class of models or structures. Suppose, however, that the class of models in question is a so-called "elementary class": i.e. that it contains precisely the models of some first-order theory T [...] Then the completeness theorem immediately yields the equivalence of Van Fraassen's account and the traditional syntactic account. (Friedman, [69, page 276-77])

And John Worrall concurs:

> The primacy of the semantic approach cannot rest on logical considerations. So far as logic is concerned, syntax and semantics go hand-in-hand –to every consistent set of first-order sentences there corresponds a non-empty set of models, and to every normal ('elementary') set of models there corresponds a consistent set of first-order sentences. (Worrall, 'An Unreal Image' [144, page 71])

Van Fraassen has answered this criticism in two stages. First he has argued, following Patrick Suppes, that first-order logic will normally be insufficient for the formalisation of any interesting theory. The Friedman-Worrall equivalence claim goes through only if a set of axioms *in the language of first order logic* exists of which the family of models is an equivalent class. The completeness theorem then guarantees the equivalence claim. But, any physical theory that appeals to the real number continuum cannot be formally

axiomatised in the language of first order logic. And virtually no interesting physical theory can make do without employing the real numbers and differential calculus.

The completeness theorem is, as Patrick Suppes[7] has emphasised, irrelevant in the formalisation of any interesting physical theories. Imagine that a scientific theory is presented in just the way that the semantic view advocates, i.e. by specifying a class of structures *directly*, without appealing to any particular language. Call this set of structures T. Suppose further, that in specifying these structures we need to mention the real number continuum. We can try to formalize T in some particular natural or artificial language L. We will write a set of axioms Ax which we can correlate with the class of semantic models in L that make the axioms Ax, and only those axioms, true. This is an equivalence class of models in L: call this set of semantic models N. However, the real number continuum is infinite, and the Lowenheim-Skolem theorem guarantees that there will be many models in N not isomorphic to any of the structures in T. So N is not a very good representation of T; and Ax is a rather poor axiomatisation of T in the language L (note, however, that no better axiomatisation of T in L is available, as the same problem recurs with any other set of axioms).

This suggests, though, a more general thought: the class of models M that constitute a scientific theory *is simply not* an elementary class –that is, there is no set of axioms (in any language) that is satisfied by all the models in the class and by only those models. Indeed Van Fraassen's second defence has consisted in stressing the distinction between structures and semantic models. He has written:

> If a theory is to be identified with the set of its models, is that set an elementary class or not? The question makes sense only if we construe "models" as referring to the models of some particular language. (Van Fraassen, [134, page 302])

And more recently:

[7] Suppes, [128, chapters 1,2].

> In a trivial sense, everything is axiomatizable, because a thing must be described in order to be discussed at all. But for logicians, 'axiomatizable' is not a vacuous term, and a scientific theory need not be axiomatizable in their sense –or as they say, the family of models may not be an elementary class. (Van Fraassen, *Laws and Symmetry* [137, page 211]).

Consider again the case of T, defined directly by delimiting its set of structures. T is not a set of semantic models; i.e. it is not a set of models in a particular language. The set N is, and it is '*a sort of image of T produced through the lens (which may be more or less distorting) of the specific chosen language*' (Van Fraassen, op.cit [134, page 302]). But moreover, as we saw, N is typically too large –it contains plenty of structures not isomorphic to any structures in T. Some unspecified subset of N, call it T^*, will contain just those structures that figure in T, accompanied by interpretations of the syntax of L. So T^* is a more perfect image of T in terms of models of L than is N. But if T^* is some unspecified subset of N, there is no guarantee that it will turn out to be an elementary class. In fact, it is now clear that the same set of structures can form an elementary class of semantic models in a language L', given some interpretation of the syntax of L', while not forming an elementary class of semantic models in a different language L, however interpreted. Unlike T^*, the theory T is not a set of semantic models that by accident fail to form an elementary class. For the elements of T, which constitute the theory, are not *semantic models*; they are instead a set of what I have called *structures*.

When a theory is presented, on the semantic view, it is not a set of semantic models that is given; rather, it is a set of *structures*. It is then possible to show that the Friedman-Worrall equivalence claim doesn't apply: structures cannot form 'elementary classes'. To define an elementary class of models we need to refer to a particular language. Semantic models would allow us to do so, as they themselves specify a language; but a structure does not specify any language. It simply makes no sense to ask of a set of *structures* whether they constitute an elementary class.

2.3 Empirical Adequacy in the Semantic Conception

Empiricists maintain that a scientific theory can be justified only by careful consideration of the relevant evidence. But in practice the assessment of a theory's empirical adequacy is a complicated affair that involves many factors other than the logical relation between different propositions. The logical positivists once championed a reduction of the theoretical to the observational vocabulary that is nowadays seen to be impossible. The semantic view of scientific theories allegedly provides an alternative. According to Van Fraassen a theory saves the phenomena by 'embedding' phenomenological structures within theoretical structures. More specifically, the phenomena must be shown to be isomorphic to some particular substructure of the theory.

2.3.1 Van Fraassen's Embedding

Van Fraassen has characterised empirical adequacy as an *embedding* relation between two different structures:

> To present a theory is to specify a family of structures, its models; and secondly, to specify certain parts of those models (the *empirical* substructures) as candidates for the direct representation of observable phenomena. The structures which can be described in experimental and measurement reports we can call *appearances*: the theory is empirically adequate if it has some model such that all appearances are isomorphic to empirical substructures of that model. (Van Fraassen, *The Scientific Image* [133, page 64])

In Van Fraassen's view theories and phenomena alike are represented by set-theoretical structures, or alternatively by sets of points in phase space. A set-theoretical structure $< D, R_i^n >$ consists of a domain D of some cardinality, and a class of relations $\{R_i^n\}$ between n-tuples of objects in the do-

main. Theories save phenomena by *embedding* phenomenological structures within theoretical structures. A phenomenological structure $P =< A, P_j >$ can be embedded in a theoretical structure $T =< B, T_i >$ (with $i \geq j$), if the former is isomorphic to an 'empirical substructure' $E =< C, T'_j >$ of the latter. The domain of the substructure is strictly a subset of the domain of the theoretical structure ($C \subseteq B$). The relations T'_j that appear in the empirical substructure are the restrictions of some of the relations T_i to the smaller domain of objects that appear in the empirical substructure ($\{T'_j\} \subseteq \{T_i|_C\}$)[8].

It is important to stress that there are not just two structures, but three: the theoretical structure, the phenomenological structure, and the appropriate 'empirical' substructure of the theoretical structure (see figure 2.1). The empirical substructure contains only relations that are already in the theoretical structure ($\{T_j\} \subseteq \{T_i\}$), restricted to the new domain ($T'_j = T_j|_C$). The theory would fail to embed the phenomenological structure if there were new relations defined on the domain of the phenomenological structure that do not have a counterpart in the theory. This is a sensible requirement: the theory could never account for all the richness and complexity of the phenomena if the phenomena were structured in ways the theory could not capture.

Hence this characterisation of empirical adequacy amounts to the following: a theory is empirically adequate *if and only if* it contains a complete representation (a 'picture') of all phenomena in its intended range[9]. If it is

[8]Relations over a domain can be defined *extensionally* by describing sets of n-tuples of objects in the domain. Thus, a one-place relation can be defined simply as a subset of the domain, a two-place relation as a set of pairs of objects chosen from the domain, and more generally an n-place relation can be defined by presenting a set of n-tuples of objects chosen from the domain. Hence, a restriction of a relation T^n, defined by a set S of n-tuples, to a smaller domain consists in taking out of S all those n-tuples that contain at least one object which is absent in the smaller domain. The restricted relation T'^n is then defined as the set of remaining n-tuples.

[9]I have here assumed that embedding is meant as a *necessary and sufficient* condition for empirical adequacy. In the passage quoted above Van Fraassen appears to present

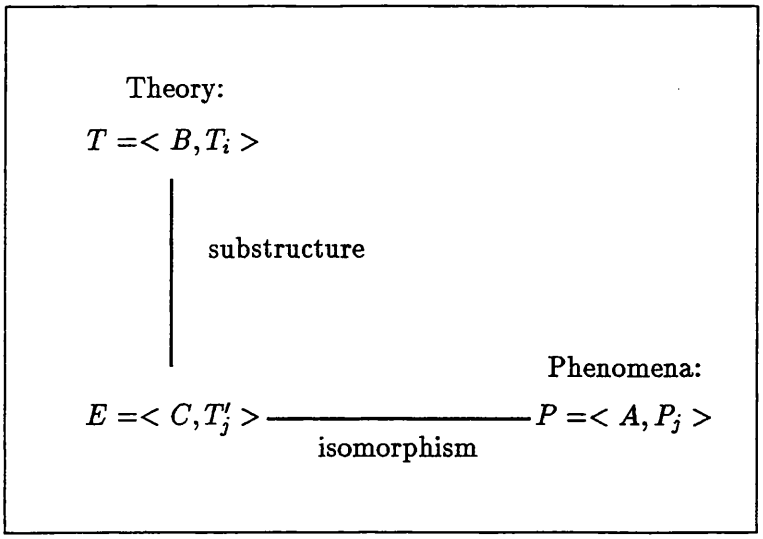


Figure 2.1: Embedding

to be empirically adequate, the theory must put forward an object to stand for each object in the domain of the phenomena, and a candidate relation for every relation that appears in the phenomenological structure. (In addition, a theory may postulate 'surplus' structure of course; a good scientific theory will typically do so.)

But the requirement of isomorphism between phenomenological structures and 'empirical' substructures has further, important, consequences. An isomorphism is a one-to-one mapping, a function that maps every element in C uniquely onto some element in A, and every relation in $\{T_j^n\}$ uniquely onto some relation in $\{P_j^n\}$. Formally, an isomorphism between two structures $P =< A, P_j^n >$ and $E =< C, T_j'^n >$ is a function $f : A \rightarrow$

embedding as a sufficient condition only. However, the *only if* part is, I think, elliptic. Embedding is commonly understood to be necessary as well as sufficient. Van Fraassen himself understands it as a necessary condition in his arguments against causal explanations of the EPR-Bell correlations –which I review in chapter 4.

C, such that if $x_1, .., x_n$ form an n-tuple of objects in A and if $y_1, .., y_n$ form an n-tuple of objects in C, then: there is some P_i^n which holds of $x_1, ..., x_n$ if and only if there is a $T_j'^n$ which holds of $f(x_1), ..., f(x_n)$; and there is some $T_j'^n$ which holds of $y_1, ..., y_n$ if and only if there is a P_i^n which holds of $f^{-1}(y_1), ..., f^{-1}(y_n)$. More concisely, an isomorphism between $P =< A, P_j >$ and $E =< C, T_j' >$ is a function $f : A \rightarrow C$ such that $P_i^n[x_1, ..., x_n]$ iff $T_j'^n[f(x_1), ..., f(x_n)]$, for any $x_1, ..., x_n \in A$; and $T_j'^n[y_1, ..., y_n]$ iff $P_i^n[f^{-1}(y_1), ..., f^{-1}(y_n)]$, for any $y_1, ..., y_n \in C$.

It follows that the domain C ($B \supseteq C$) that the theory must put forward as candidate for representing the phenomena must be of the same cardinality as the domain of objects A in the phenomenological structure. The number of relations T_j' defined over this domain must be identical to the number of relations P_j that appear in the phenomenological structure. And, although the objects and their properties (represented by the n-place relations P_i^n on A) need not be preserved with the mapping, the properties of the *relations* that denote those properties must be so preserved[10]. The existence of an isomorphism between two structures is, so to speak, a statement that the *structures* have identical properties. In the context of the semantic view, isomorphism *is* structural identity.

There is one fundamental difference between this characterisation of empirical adequacy in the semantic view, and the old syntactic criterion of

[10]The relation T_j' put forward by the theory to match the relation P_j in the phenomenological structure will not hold of the *same* objects because the phenomenological structure is not *part of* the theoretical structure. As a relation can always be defined extensionally by appeal to sets of n-tuples of objects in its domain, and as the domains of the phenomenological and the theoretical structure will typically be different, it follows that the relations themselves will be different. In other words, isomorphism between structures does not require that the *objects* in the structure be the same, nor that the *properties* of those objects be identical. However, if there is to be an isomorphism between the structures the domains A and C must be of equal cardinality, and the corresponding relations T_j' and P_j must possess identical properties. For instance, an equivalence relation over a domain of cardinality x is mapped onto another equivalence relation over a domain of equal cardinality, etc.

empirical adequacy. It becomes possible, in the semantic view, to isolate the empirical import of a theory: it is the set of substructures the theory puts forward as candidates for the representation of phenomena. There is no problem in isolating the *observable* content of the theory in this manner either. The theoretical domain can be restricted, if so wished, so as to yield an empirical substructure with a domain of observable objects only. We would then be interested in the properties that the theory ascribes to these objects, so we would look for a restriction of the properties postulated by the theory to those objects. There is of course no reason to think that these properties themselves will be observable, nor is it required in the 'embedding' conception that they should be so (for more on this, though, see footnote 11). And there certainly is no reason why we should have to refer to these objects and relations, when we speak about them, in a purely 'observational' language. Such requirement would only follow if the models that, on the semantic view of theories, constitute a scientific theory were *semantic models*. For in that case, as Michael Friedman makes clear:

> the empirical substructures in question are definable in the language of [the theory...] Consider one of Van Fraassen's empirical substructures: a set of objects (the observable objects) together with appropriate properties and relations. Now look at the terms the theory uses to denote these properties and relations: these comprise our observational vocabulary. Look at the sentences built up from these terms by truth functions and quantifiers restricted to the set of observable objects: these will be just the observational sentences. T will be empirically adequate in Van Fraassen's sense just in case its (syntactic) consequences within this class of observational consequences are true. (Friedman [69, page 277])

If the class of models that constituted the theory was an 'elementary class', there would be a syntactic representation of the theory T, and an equally syntactic representation of the empirical content of this theory. In rejecting

the construal of models as *semantic* models, we have preempted this possibility. A theory is a collection of models; fortunately these are not semantic models; hence they cannot be said to form an elementary class; so there need be no corresponding syntactic characterisation of any of the features of a scientific theory, and that includes its empirical adequacy.

2.3.2 Friedman's Model-Submodel Reduction

Embedding nicely captures an empiricist intuition about confirmation: the degree of confirmation of a theory does not measure the likelihood that the theory is true but, rather, the likelihood that the theory is empirically adequate. According to the embedding conception a theory provides *representations* of the phenomena. The theory provides structures, and it delineates those substructures that are intended to directly model the phenomena. The theoretical structures are constituted by certain mathematical or physical entities (such as numbers and n-tuples of numbers, lines, planes, manifolds, vectors and n-tuples of vectors; electrons, atoms, molecules, etc.,) and some relations defined over them (operations such as norm, angle, scalar product, trace; the relations of spacelike, lightlike and timelike separation in relativity theory; and physical properties such as charge, mass, etc). The entities and relations in the theory are emphatically *not* to be thought of as existing in reality. The mathematical structures are not *real*; rather they are mere candidates for *representations* of the phenomena. Theories may be more or less successful in this task of representing, of course, and a theory's accumulated degree of confirmation measures just the overall success of a theory in successfully representing phenomena. For convenience, I will refer to this picture of the relation between theories and phenomena as the 'representational' picture.

Van Fraassen's constructive empiricism goes even further, by restricting the phenomenological structures to sets of observable entities and their properties. The constructive empiricist advises us to suspend belief in the existence of the unobservable entities and processes postulated by scientific theories. We may choose to accept a scientific theory, as a research pro-

gramme, and we may have confidence in its future predictions, but the only *belief* that is involved in such acceptance is the belief that the theory is empirically adequate, not the belief that it is true:

> Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate. (Van Fraassen, *The Scientific Image* [133, page 12])

Van Fraassen's 'selective' epistemology implies that it is always possible to identify the empirical import of a theory, and to separate it from the 'surplus' theoretical commitment. The semantic view, in its representational version, provides a schema: the empirical content of a theory in constituted by those empirical substructures that the theory makes available to model the phenomena. We are then advised to believe only in the existence of those entities and processes that appear in the phenomenological structure (i.e., according to the constructive empiricist, those entities that are directly observable); and we are told that we are not required to believe in the existence of the unobservable entities that the theory postulates.

Michael Friedman, on behalf of the scientific realist, has put forward an argument against the representational picture. In his view, a more proper account of the relation between theory and phenomena (particularly in the context of space-time theories) is rendered by the notion of model-submodel reduction. In contrast to an embedding, a reduction requires only two structures: the theoretical structure and the phenomenological structure. A theory is empirically adequate if it can literally subsume the phenomena (see figure 2.2). In a reduction the domain of the phenomenological structure is *a subset* of the domain of the theoretical structure ($A = C \subseteq B$), and the properties found in the phenomena are precisely the same relations that the theory postulates, restricted to the domain of phenomenological entities and processes ($\{P_j\} = \{T'_j\} = \{T_j|_C\}$). As Friedman writes:

> Under this construal T functions as a genuine explanation or *reduction* of the properties of P, for elements of P are literally

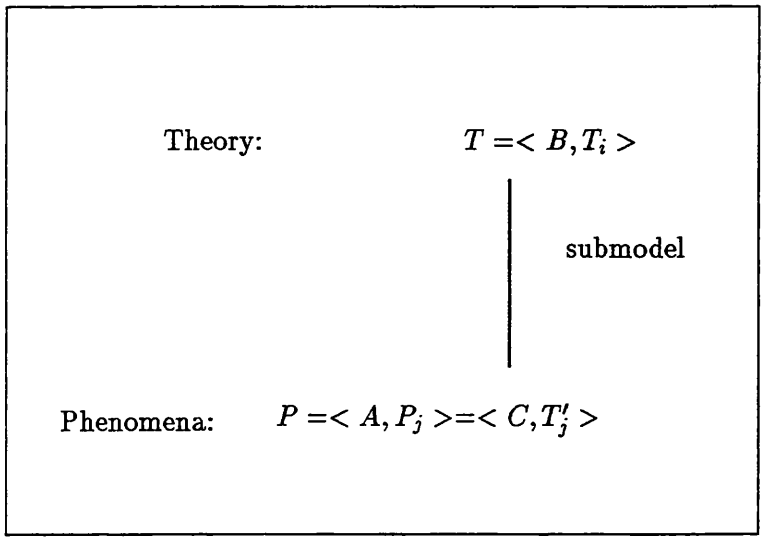


Figure 2.2: Model-Submodel Reduction

identified with elements of T. (Friedman, *Foundations of Space-Time Theories* [70, page 236].)

As a consequence, on this alternative 'reduction' picture, if a theory is shown to account for some new phenomenon, this constitutes a reason to believe that the whole theory is true –including the 'surplus' theoretical relations $\{T_{j+w}\}$ (with $w \leq i-j$), and the 'surplus' domain $B-C$. For it would make no sense to claim, on this view, that the phenomenological structure is true but the theoretical structure is not, when the phenomenological structure is actually a part (a '*chunk*') of the theoretical structure. Therefore, in the reduction picture, degree of confirmation is a measure of the likelihood that the theory is true.

The main argument in favour of constructive empiricism (and its close ally in the philosophy of space and time, relationalism) is that a scientific theory, considered in isolation, has exactly the same empirical consequences on the representational picture as on the reduction picture. That is, if we

consider the theory in isolation from other theories, and focus exclusively upon its relation with the phenomena for which it is intended, the theory receives exactly the same degree of confirmation on the representational picture as on the reduction picture. As the representational picture is committed to the existence of fewer entities, refraining as it does from reifying any theoretical entities, it is always preferable on grounds of ontological parsimony.

An initial objection to the representational picture is that it is not possible to make do without theoretical 'surplus' in the description of the phenomena. This 'surplus' will often induce theoretical properties and relations on objects in the phenomenological domain which are normally required to state accurate laws about the behaviour of these objects. But, the objection continues, the representational picture implies that no 'surplus' can ever be employed in the description of the phenomena. Asserting that the phenomena $P =< A, P_j^n >$ is merely embeddable in a theory $T =< B, T_i^n >$ will not induce the necessary theoretical properties and relations. If T_v^1 $(i > v > j)$ is one such theoretical relation, and unless T_v^1 is definable from the set of phenomenological properties $\{T_j'^n\}$, there will be automorphisms of the domain of the empirical substructure C, that will leave the isomorphism with the phenomenological structure entirely intact while crucially altering the extension of T_v^1 in the empirical substructure. In other words, there will be two embeddings f and g of P into T that take every P_j^n into its corresponding $T_j'^n$ and yet such that, for some object $a \in A$, $T_v^1(f(a))$ and $\neg T_v^1(g(a))$. The objection, however, is not a very good one. For there is no reason why the representationalist cannot simply expand the empirical substructure by restricting this new relation T_v^1 to the domain C. In fact, given some justification, the representationalist could introduce *all* 'surplus' theoretical relations $\{T_{j+w}^n\}$ (with $w \leq i-j$), appropriately restricted to C, into the empirical substructure[11].

[11] Admittedly, however, this argument may cut some way against the constructive empiricist. For the introduction of T_v^1 into the empirical substructure entails the existence of a corresponding property in the phenomenological structure, under the assumption that both structures continue to be isomorphic. Is the constructive empiricist happy to intro-

Friedman, who is well aware that a different strategy is needed to defend the reduction picture, has proposed to block the main premise of the above argument in favour of the representational picture. He denies that in all occasions a scientific theory will gain the same amount of confirmation on the representational picture as on the reduction picture. If we considered the relations between scientific theories, we could see how the representationalist and the reductionist accounts may ascribe different degrees of confirmation to the very same theory. In particular, Friedman claims, in cases of theory-unification there is a substantial boost in confirmation for the individual theories that make up the unified theory whenever the unified theory comes up with a successful new prediction in a different domain. The reduction picture can accommodate this fact naturally, while the representational picture is unable to do so.

An interesting reply to Friedman has been given by Morrison [103]. However, this debate, which is fascinating in its own right, is not directly relevant to my concerns in this Thesis. The reductionist and the representationalist can equally account for the theoretical induction of properties and relations into the phenomena; they agree on what the degree of confirmation of an individual theory is, whenever taken in isolation, in the presence of a phenomenon; and, more importantly, they agree that a theory always picks up confirmation from phenomena that the theory accounts for (although, naturally, they disagree as to whether increased confirmation indicates that

---

duce *theoretical* relations among the *appearances*, if it is in the restricted form in which they are taken to apply to observable entities only? (I am indebted to Hasok Chang for raising this question). I am not sure. Perhaps Van Fraassen's *internalism* can help here. For *internalism* –the thesis that it is up to science itself to decide what to count as observable and what not– enables us to bring into phenomenological structures relations that were previously thought to be unobservable. (N.B. *Internalism* is built into Van Fraassen's definition of a scientific theory: "to present a theory is to specify a family of structures, its models; *and secondly*, to specify certain parts of those models (the empirical substructures) as candidates for the direct representation of observable phenomena").

In any case, the representationalist is in no trouble at all here –for he is not committed to the view that the entities and relations in the phenomena should be 'observable'.

the theory is more likely to be true, or just more likely to be empirically adequate).

Moreover they agree on what the logical relation is between their own positions. The representationalist thinks that there may be many different ways of mapping a phenomenological structure onto empirical substructures of a theory (in other words there may be different mappings: f and g above are an example), while the reductionist thinks that there is only one possible mapping, namely identity. Hence, they both agree that reduction is logically stronger than embedding. And this is important because it shows that if a theory is unable to *embed* a particular phenomenon then it cannot be empirically adequate *on either characterisation of what it means for a theory to be empirically adequate.* Such a theory, therefore, cannot pick any confirmation at all from that particular phenomenon.

2.4 The Empirical Basis of Science

I have been discussing different conceptions of scientific theories, with special emphasis on the relation between theory and phenomena. Van Fraassen and Friedman have established the requirements that this relation must obey if the theory is to receive confirmation, on an empiricist and on a realist construal. Their work is heavily informed by a conception of theories as structures. But, so far, I have had little to say about the other *relata* of the confirmatory relation, namely *phenomena.* A more detailed discussion of what constitutes phenomena is imperative, though, for Van Fraassen's and Friedman's epistemological purposes. Theories are to receive confirmation from the phenomena they account for; but unless we have good grounds for believing that the phenomena are true, Van Fraassen's and Friedman's strictures will count for nothing in persuading us to believe in the truth of our theories. In other words, phenomena must constitute a solid epistemic basis for scientific theories.

Both Van Fraassen and Friedman assume that it is possible to describe a phenomenon as a structure. This is of course, absolutely essential for

their proposals to work, as both embedding and reduction are defined as relations between *structures*. Van Fraassen mentions that he first saw this kind of formalised structures of the phenomena in the works of Wojcicki [142], Przelecki [112], Dalla Chiara and Toraldo di Francia [41] and Suppes [126]. I shall concentrate on Suppes's classic paper 'Models of Data' [126].

2.4.1 Models of Data

Suppes argues that there is a hierarchy of set-theoretic models all the way down from theory to experimental data. He concentrates on statistical theories, where the basic hierarchy contains five layers: models of theory, models of experiment, models of data, methods and models of experimental design, and *ceteris paribus* conditions on extraneous factors in the experiments actually performed. Suppes illustrates the former three by means of an example from learning theory. We are investigating the responses exhibited by some organisms to an stimulus, subject to patterns of reinforcement on their responses. We assume that there are only two possible responses, A_1 or A_2, and after each response the individual receives one of two possible reinforcements, E_1 or E_2. An experimental trial consists of a finite number N of sequences each formed by a finite number m of observed pairs (A_i, E_j), where $i, j = 1, 2$. The first member of the first pair, A_{i1}, represents an individual's first response to the stimulus; the second member of the first pair, E_{j1}, represents the reinforcement immediately impressed upon the individual. The stimulus is then repeated; the first member of the second pair, A_{i2}, represents the response of the individual to this repeated stimulus; E_{j2} represents the subsequent reinforcement; and so on (there are N organisms, and the stimulus is presented m times to each one).

The theory involved is a linear response theory. Suppes takes a model of this theory to be a structure $S =< X, P, \Theta >$, where X is the (infinite) set of all possible (infinite) sequences of pairs A_i, E_j, and P is a probability function defined over the smallest Borel field that contains the field of cylinder sets of X (roughly: the smallest field generated by the elements in X over which it is possible to define a Lebesgue probability measure),

Θ is a learning parameter, not observable and hence "*not determinable by experiment*" [126, page 26]. By forming equivalence classes x_n of sequences identical through a particular trial $n \leq m$ we can impose on the probability measure P constraints required for a linear response theory. These conditions need not concern us here; it suffices to say that they are conditions on the relative frequencies of reinforcements E_j conditional on responses A_i.

Suppes then constructs a 'model of the experiment' $< Y, P' >$ where Y is a finite set, consisting of all possible sequences formed by a number m of pairs (A_{iw}, E_{jw}) $(w \leq m)$, and P' is a probability function on the field generated by the power set of Y. In the experiment the stimulus is presented m times. So, correspondingly, the elements of Y are all the possible sequences of length m. This model is applicable only if the probability measure P' is defined to give relative frequencies of reinforcements on responses, $P'(E_i/A_j$ in every trial that agree with the probability distribution P in the theory.

A model of experiment is not a data-model; it is a truncation of the theory to the conditions in the experiment. Y contains 4^m possible sequences. But the data will at most amount to a number N of such sequences, where N is the number of organisms on which the experiment has been performed; hence a much smaller number. A 'model of the data', or data-model, is defined by Suppes as an N-tuple of sequences in Y where N is the number of sequences of experimental outcomes, or sequences of pairs (A_{iw}, E_{jw}) $(w \leq m)$ of an actual experimental run, and where the conditional relative frequencies of reinforcements $\frac{n(E_i)}{n(E_j)}$ approximately fit the distribution induced by the probability measure in the model of experiment:

> An N-tuple realization is a model of the data if the conditional relative frequencies of E_1 and E_2 reinforcements fit closely enough the probability measure P' of the model of the experiment. (Suppes, op.cit. [126, page 31])

I go through these details because I want to emphasise that, in Suppes' view, the relation between theory and data is a tremendously complex one. The theory contains continuous distributions but the data is finite, and one has to figure out how well the relative frequencies in the finite sample agree

with the distribution function. The theory has to be 'prepared' to meet the demands of the experiment, by truncating the continuous distribution P over the infinite set X into a discrete measure over the finite set Y. The data has to be similarly 'prepared' in a form that the theory is able to account for, by selecting those runs whose frequencies of reinforcements fit at least approximately the conditional probabilities in the model of the experiment:

> The central idea, corresponding well, I think, to a rough but generally clear distinction made by experimenters and statisticians, is to restrict models of data to those aspects of the experiment which have a parametric analogue in the theory. A model of the data is designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory. (Suppes, op.cit. [126, page 31])

In practice the decisions taken to 'prepare' the theory and the data for each other, –by imposing certain constraints on the probability distributions in the theory, or by selecting a subset of experimental runs–, can only be made on a case-by-case basis. There are no general rules. These judgements demand practical skills and a fine sense for the details of the experiment; they demand a considerable amount of *tacit* knowledge and a careful consideration of the particular conditions under which the experiment is performed. Here I have not dealt with the sort of detailed considerations that play a role in Suppes' example. (They are many and subtle; Suppes himself refrains from discussing all of them in detail.) The main point that I want to emphasise, following Suppes, is that theories do not confront raw data, but rather, specifically tailored, contrived, and highly conceptualised 'data-models'.

2.4.2 The Empirical Basis: Data or Phenomena?

The same point has been recently reinforced and expanded in an interesting way by James Woodward and Jim Bogen [14], [143], who draw the distinction between data and phenomena. Facts about data describe sets of data-points on a graph, records of scintillation spots on a fluorescent screen,

registered tracks in a bubble-chamber, etc. Data live in the laboratory; they are highly dependent upon the particular experimental context in which they are observed and recorded; and they are generally short-lived: as better, more accurate experiments are performed, the data from previous experiments tends to get ignored, or reinterpreted. By contrast, phenomenological facts describe well-established general patterns in nature, and effects. For instance, the fact that *metals dilate in the presence of heat*, is a rather typical phenomenological fact. Phenomena constitute firmly established bodies of scientific knowledge; they are independent of the particular conditions under which experiments are performed; and, besides, they are relatively theory-independent, in the sense that they tend to 'last long': facts about phenomena are manifestly impervious to theory-change.

Woodward and Bogen's main claim is that phenomena, not data, constitute the empirical basis of science. First, Woodward and Bogen argue that scientific theories are advanced by scientists as explanations of phenomena, but are not normally intended as explanations of data: "*typically, scientific theories are expected to provide systematic explanations of facts about phenomena rather than facts about data*" (Bogen and Woodward, 'Saving the Phenomena' [14, page 322]). Theories must account for (i.e. be empirically adequate of) phenomena; but they are not required to account for any data. Secondly, Woodward and Bogen argue, what is observed and gets recorded during an experiment is data, not phenomena. Phenomena are instead *inferred* from bodies of data: "*Data, which play the role of evidence for the existence of phenomena, for the most part can be straightforwardly observed. By contrast [...] phenomena are detected through the use of data, but in most cases are not observable in any interesting sense of that term*" (op.cit. [14, page 306]).

Woodward and Bogen's target is the requirement, defended by some logical positivists and by Van Fraassen's constructive empiricist, that scientific theories account for *observable* phenomena. In Woodward and Bogen's view the expression 'observable phenomena' is a category mistake. As they write: "*phenomena for the most part cannot be observed and cannot be reported by*

observational claims" (op.cit. [14, page 343]). Consider, ask Woodward and Bogen, the well-established phenomenological fact that 'lead melts at 327 degrees C'. In what way is this fact ascertainable by observation? Is it an observable fact? Ernest Nagel writes:

> The law that when water in an open container is heated it eventually evaporates is a law which formulates a relationship between observables, and so is the law that lead melts at 327 degrees Celsius. (Nagel, *The Structure of Science* [107, page 79])

But according to Woodward and Bogen this phenomenological claim is not observed, nor is it ascertained directly by observation. Instead it is *inferred* from observed data:

> Despite what Nagel's remarks suggest one does not determine the melting point of lead by observing the result of a single thermometer reading. To determine the melting point one must make a series of measurements. These constitute data. [...] What we observe are the various particular thermometer readings –the scatter of individual data points. The mean of these [the estimate for the melting point] does not represent a property of any particular data-point. [...] So while the melting point is certainly inferred from data, on the basis of a theory of statistical inference, the sentence 'lead melts at 327 +/- 0.1 degrees' does not literally describe what is perceived or observed. (Bogen and Woodward, op. cit. [14, page 308])

An important part of Woodward and Bogen's argument concerns their defence of the claim that the assessment of the reliability of data does not require a scientific theory to provide explanations of the data. They provide a number of case studies to illustrate and defend this claim. Consider again the melting point of lead. A number of measurements are made on samples of different sizes, under different experimental conditions, in different laboratories. A thermometer sensor is fixed to a single sample of lead, and

a reading is made and recorded as soon as melting begins. The sensor is then detached, and the sample is cooled down, before another experiment is made on the very same sample. These recorded data-points constitute *evidence* for or against the phenomenological claim.

It is striking, however, that the precise melting point may correspond to no actual data-point at all. Many data-points will lie very close to the mean. But some other isolated data-points will lie far away from the mean. Entire sets of data-points belonging to a particular sample may lie consistently far away from the mean –thus suggesting that something went wrong with the entire trial on one particular sample. Sometimes it will be impossible to record any values –perhaps the sensor did not work appropriately, or the thermometer failed. On occasions the whole set of data-points for a particular sample will exhibit temporal correlation –perhaps an indication that the thermometer wasn't brought back to zero on repetition of each new experiment, or that the sample of lead did not cool down sufficiently. And so on. Many measurement outcomes will need to be discarded as they result in unreliable data. But, in order to assess the reliability of the data, we don't need to provide a comprehensive explanatory account of what went wrong in each discarded case[12]. We may rest content with considerations such as "*whether the data are replicable, whether various confounding factors and other sources of possible systematic error have been adequately controlled, on statistical arguments of various kinds, and on one's procedures for the analysis and reduction of data*" (Bogen and Woodward, op.cit. [14, 327]).

Woodward and Bogen's account of phenomena has strong similarities

[12]There would be a very large number of factors that would need to be taken into account in each explanatory instance; and such a complete theoretical explanation may not be available. But, in fact, a complete explanatory account is actually undesirable. As Bogen and Woodward point out, if the determination of the reliability of data as evidence for various phenomena depended upon our possessing a general and comprehensive theoretical account of the causal mechanisms which produce the data, claims about phenomena would be far more fragile than they actually are. Phenomena would not be impervious to theory-change in the way they manifestly are.

with Suppes' account of data-models, but there is one important difference. For Suppes data-models are primarily employed to test scientific theories. This is indeed the reason why the data-model must be 'prepared' in the light of the theory's assumptions and requirements –so that the data can be put to use as a meaningful test of the theory. By contrast, *models of the phenomena* are often constructed to describe stable phenomena or to elucidate them, and not necessarily to test theory –although they may be constructed with the help of theory.

According to Woodward and Bogen a phenomenon is inferred from data, and not always with the view of testing any particular theory. However, as I have noted, when a theory is tested, according to Woodward and Bogen, it is normally tested against well established phenomena. So there is a two-stage process. First a phenomenon is inferred and some description of it is given in a model. Second, a theory is proposed and it is tested against the accepted phenomena in its field. A model of the phenomena, as I want to describe it, is part of the first stage of the process, and is hence typically independent of the testing of theories[13]. One example is the London theory of superconductivity that I described in the first chapter. The Londons' aim was not to test electromagnetic theory, but to refine our understanding of superconductivity phenomena. Their theory constituted a model of the phenomenon of superconductivity –among other things it predicted the Meissner effect. The claim that superconductors expel magnetic flux is surely a phenomenological claim, if any claim ever is: it is quite impervious to theory change, and its truth does not depend upon any of the conditions in the experiments performed to verify it. In contrast, a data-model, which may record the average experimental rates of expulsion of flux in the Meissner-Ochsenfeld experiment, is a highly contextual model: it may be true of the data collected in that particular experiment; but it cannot constitute a model of the

[13]Models of the phenomena must not be confused with the kind of *mediating models* that I discussed in the first chapter. Models of phenomena should also be distinguished from the *phenomenological models* that I considered briefly in my response to French and Ladyman, in section 1.4.4. The London treatment of superconductivity is *both* a model of the phenomena, *and* a phenomenological model.

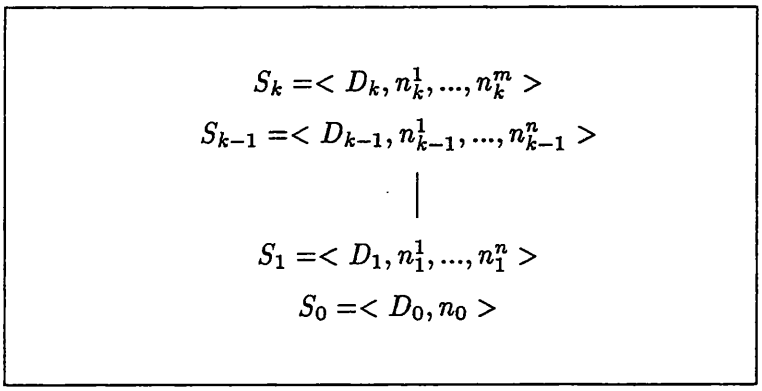


Figure 2.3: Layers of Structures

Meissner effect *per se.*

Mathias Kaiser [93] has proposed an account in terms of hierarchies of structures. In his picture a *model of a phenomenon* S_k sits at the top of a hierarchy of *models of data* (see figure 2.3). According to Kaiser, the bottom layer's structure has as its domain only objects that can be directly observed, and the only relation allowed, D_O, is ostension[14]. The next structure up is defined by introducing a set of new relations over a perhaps different domain. The whole hierarchy is defined in a number of steps, or layers, each of which corresponds to a change of elements in the domain, or a change in the set of relations. These operations are intended to represent

[14]In another, puzzling, paragraph Kaiser asserts that the objects in the bottom layer are *all those things that are to be subjected to scientific scrutiny, i.e. that are to be measured, weighed, radiated, dissoluted, accelerated, etc.*" [93, page 125] But it is obviously not the case that everything measurable must be definable by ostension, or even observable. As Mary Morgan has pointed out to me, many economic entities (inflation, unemployment, bullion, etc.) can be measured very accurately, yet can never truly be observed. In this respect physics is actually not that different; consider for instance measurements of electric currents in circuits by means of amperometers, or measurements of field strengths by induction techniques. I take it that Kaiser's conflation is a result of a too narrow focus on his case studies in paleomagnetism.

scientific practice, hence they correspond to operations and redefinitions of the domain that are actually performed in the course of scientific research. The passage from one structure to the next one up is grounded upon what Kaiser calls '*inference tickets*' which come in two kinds: as warrant for a redescription of the domain, or as legitimation of new types of operations and computations. Theoretical knowledge is involved –although not necessarily knowledge derived from one and the same theory. 'Inference tickets' are provided by probing techniques and instruments, error theory, test-methods to eliminate useless samples, statistical techniques for the treatment of data, etc.

Kaiser's hierarchy of models of data has three interesting features. First, it represents a model of the phenomena as a structure $S_k =< D_k, n_k^i >$, as required by the advocates of the semantic view. Second, it allows for the fact that the construction of models of phenomena may be constrained by (although not dictated by) theory, as in Suppes's example. But it also makes clear that the construction of a model of phenomena usually is independent of the testing of any particular theory. Finally, Kaiser shows that an adequate scientific theory ought to account for the phenomena, which is represented by the top structure in the hierarchy of structures.

As a consequence, however, it is necessary to forgo the requirement that an empirically adequate theory must 'embed' or 'reduce' *data*: on Kaiser's proposal the model of the phenomena will normally fail to embed or reduce the lower models of the hierarchy. The changes in the domain and in the set of relations are not necessarily driven by one unique body of knowledge as in Suppes' example, so there is no guarantee that the domains and relations of structures in different layers will be appropriately 'nested'. Hence, an empirically adequate theory (a theory that accounts for the phenomena) will typically not embed or reduce the data that serves as evidence for the phenomena. On Woodward and Bogen's view, however, this is as it should be: data rarely constitute evidence for a scientific theory; instead the evidence for or against theories is typically to be found in well-established phenomenological facts.

2.5 The Application of Scientific Theories

Ian Hacking[15] denies that scientific models are generally 'doubly' models, i.e. models of phenomena as well as models of theory. Some examples of models of phenomena include early models for continental drift phenomena, the early models of the atom (the Bohr atom), the billiard ball model in phenomenological thermodynamics, models for stellar structure, gravitational lensing models. How do these models relate to models of theory: models of the continental drift theory, the hydrogen atom in quantum theory, models of the kinetic theory, models in the quantum theory of radiation, and the Friedmann models in cosmology? According to the advocates of the semantic view if a theory is to be empirically adequate its models must embed, or reduce, all models of phenomena in the domain of the theory. Can this account of empirical adequacy be used to describe the *application* of scientific theories? Does the domain of application of a theory coincide with its domain of empirical adequacy?

2.5.1 Revisiting the London Account

In section 3 I showed that 'embedding' is a logically weaker constraint than reduction. A theory may fail to reduce phenomena, and yet a part of the theory may be shown to be isomorphic (structurally identical) the phenomena. In this sense, embedding is a more generous constraint than the alternative realist, characterisation of empirical adequacy as reduction. But it is not an *empty* constraint: some models of phenomena are not embeddable in theory. A model of the phenomena that contains relations that have no structural counterpart in physical theory has no means to be embedded. Some models of the phenomena cannot be embedded in *any* theory, not even in those theories employed to build the models in the first place. Some examples include gravitational lensing models in cosmology, stellar structure models in astrophysics, and models of SQUIDS (Superconducting Quantum Interference

[15] Hacking [80, page 216].

Devices).

The model that Fritz and Heinz London gave of superconductivity phenomena is another instance. Let me now review the London model from the point of view of the semantic view. In this case, there are three structures: the theory of classical electromagnetism, which I take to include at least Maxwell's equations; the 'acceleration equation' model of Becker, Sauter and Heller which entails the following relation for the magnetic flux:

$$\Lambda c^2 \nabla^2 (\mathbf{H} - \mathbf{H_0}) = \mathbf{H} - \mathbf{H_0} \tag{2.5.1}$$

where $\mathbf{H_0}$ denotes the magnetic field at the time t=0; and the London equation which models the Meissner effect by imposing a different constraint on the dynamics of the magnetic flux:

$$\Lambda c^2 \nabla^2 \mathbf{H} = \mathbf{H} \tag{2.5.2}$$

In section 5.2. in the first chapter, I gave two alternative pictures of the state of classical electromagnetic theory at the time of Meissner's discovery. First, I argued, it is possible to take 'electromagnetic theory' in a historical sense, as constituted by all applications to electromagnetic phenomena known at the time. On this picture, the 'acceleration equation' model is part of classical electromagnetic theory, and the following identification seems irresistible: the theory (classical electromagnetism) contains an empirical substructure (the 'acceleration' model of Becker, Sauter and Heller) which it puts forward as a candidate for representing the phenomena (the Meissner effect). As the Meissner effect is accurately described by the Londons' model, we can take the latter to provide its phenomenological structure. This may of course not be the only possible identification. But let me consider first whether, with this identification in mind, electromagnetism can be said to account for the phenomena (in the sense of being empirically adequate).

We are looking for an isomorphism between the phenomenological structure and the empirical substructure. The London model is close to being isomorphic to the 'acceleration' model. First, the two domains are isomorphic: for every physical entity ($\mathbf{j}$, $\mathbf{H}$, etc) in the domain over which the

'acceleration' model is defined, there is a corresponding entity in the domain of the London model. Second, at least one relation over the domain is isomorphic, namely the relation that accounts for a constant current in the absence of external fields, expressed in the equation:

$$\nabla\Lambda\frac{d\mathbf{j}}{dt} = -\frac{1}{c}\frac{d\mathbf{H}}{dt} \tag{2.5.3}$$

As I mentioned in the first chapter of the Thesis (section 1.4.4) this equation represents the common part to the London model and the acceleration model.

Nevertheless there fails to be a complete isomorphism. The two models are not structurally identical in the required sense. Equation (2.5.2), which entails the '*fundamental law*' of superconductivity, yields a prediction that the 'acceleration equation' model cannot match. In the London model we find a relation, lacking in the 'acceleration' model, that establishes that the magnetic flux after the phase transition is zero, regardless of what the flux is before the transition. On the other hand, according to the 'acceleration' model there should be some nonvanishing flux after the transition. The phenomenological structure contains one relation –expressed in equation (2.5.2)– that is not in the theory, while the theory contains one relation –equation (2.5.1)– that is not in the phenomena. The relation in the theory is an equivalence relation, characterised by an identity mapping that leaves the magnetic flux invariant, while the relation in the model is not even a symmetric relation (at least not when there is some initial magnetic flux).

There is a caveat, though: the fact that we have failed to find an empirical substructure of the theory isomorphic to the phenomena does not necessarily mean that the theory does not have one. We may have misidentified the correct empirical substructure of the theory: there could be another substructure, unknown to us, that is totally isomorphic to the London model. And yet, the 'acceleration' model is *inconsistent* with the London model. So if the unknown substructure –the 'good' one– is to account for the Meissner effect, it must contain at least one relation inconsistent with the 'acceleration' model. This generates a paradox: the theory contains

two substructures, intended for the same phenomenon, that yield mutually inconsistent predictions. A theory, if consistent, cannot yield inconsistent predictions *for the very same physical phenomenon.* Hence, there must be at most one substructure of the theory that applies to superconductivity and, given the historical construal of electromagnetic theory, this must be the 'acceleration' model –which fails to be isomorphic to the Meissner effect.

To sum up, on Van Fraassen's characterisation of empirical adequacy, –and *a fortiori* on Friedman's too–, classical electromagnetism, if construed in this historical sense, is not empirically adequate (of superconductivity phenomena). This cannot be very surprising –for we did not expect the acceleration equation theory to account for the phenomena in the first place[16]. But I hasten to say that this is not a conclusive result. For there is an alternative picture of electromagnetic theory, which takes classical electromag-

[16]I should mention here the *partial structures* approach developed by Newton Da Costa and Steven French [38]. A *partial structure* S is a structure $< D, R_i >$, where some of the R_i are *partial* relations. A partial n-place relation R over a domain D is a triple $< R_1, R_2, R_3 >$ where R_1 is the set of n-tuples that satisfy R, R_2 is the set that don't satisfy R, and R_3 is the set for which it is not known if they satisfy R.

In response to an earlier draft of this chapter, Otávio Bueno [20] has suggested that it may be possible to lay out a *partial*, if not a total, isomorphism between the London model and the acceleration model. (A partial isomorphism between two partial structures $P =< A, P_j >$ and $E =< C, T'_j >$, is a mapping $f : A \rightarrow C$ that preserves the properties of all the partial relations $\{P_j, T'_j\}$). I have doubts that Bueno's approach will work. Consider an n-place relation $P_1 =< P_{11}, P_{12}, P_{13} >$: no n-tuple of elements of A can figure simultaneously in two of the sets defined by the extensions of P_{11},P_{21} or P_{31}. Hence, a partial isomorphism between P and E reduces to a set of three isomorphic mappings which take each of P_{1i}, P_{2i}, P_{3i} into the corresponding $T'_{1i}, T'_{2i}, T'_{3i}$. The problem, as I see it, is that the London model contains a relation that is explicitly denied by the acceleration model, and viceversa. So there is no room to exploit the main insight of the partial structures approach, namely that some n-tuples may belong neither to the extension of a relation, nor to the extension of its negation.

Let me emphasise that this result does not invalidate in any way the *partial structures* approach to scientific theories. On the contrary, if the formal notion of partial isomorphism is to provide an extended characterisation of empirical adequacy, we should hope that it fails here, for we have independent reasons to believe that in this case the theory, as understood in the extended historical sense, is not empirically adequate.

netism to be constituted by the deductive closure of Maxwell's equations. However, neither the acceleration equation nor the London model belong to this closure. On this abstract picture, the theory does not 'reach far enough'. It is neither confirmed nor disconfirmed by the Meissner effect. It lacks empirical substructures capable of representing superconductivity phenomena –the phenomena simply lie outside the theory's domain of empirical adequacy.

Hence, on neither picture is the theory of classical electromagnetism empirically adequate of superconductivity. But note that the failure of empirical adequacy does not seem to impugn the claim that the London model of superconductivity was a very successful application of classical electromagnetic theory. The model accommodated both defining features of superconducting behaviour, and it had predictive power: it yielded accurate rates of expulsion of flux for the correct temperatures and values of the external magnetic field, and it accounted for the current and fields in superconducting spheres and wires accurately. Bardeen, for instance, writing nearly 20 years after the advent of the London model, notes that:

> The London theory gives a complete electrodynamics of superconductors which has been very successful in correlating and predicting results of experiments. (Bardeen, op.cit. [9, page 284])

In Duhem's words the phenomena, although not embedded, were, in a sense, 'saved' in electromagnetic theory.

2.5.2 Instrumental Reliability

In the first chapter (section 5.3) I introduced the notion of *instrumental reliability*, and I argued that this notion should be distinguished very carefully from the more common notion of empirical adequacy. By now I hope I have made clear why. I want to link the instrumental reliability of a theory to its effectiveness as an tool in application. However, a theory's domain of application is typically much larger than its domain of empirical adequacy: if

the notion of instrumental reliability is to capture the various ways in which scientific theories are applied, it must be kept separate from the notion of empirical adequacy.

Similarly the explanatory power of a theory does not constitute a guide to its instrumental reliability –a theory can be employed to generate a model of phenomena that it cannot provide explanations for. Nancy Cartwright [29] has argued for a separation between the explanatory power of a theory and its descriptive power. She argues that in many occasions theories provide descriptions of phenomena that they cannot explain. My concerns to separate application from confirmation in many ways mirror Cartwright's concerns to separate application from explanation. She writes

> This is how we solve problems in mathematical physics. Some of the techniques are a standing part of the theory; but others are *ad hoc*, to the problem at hand. Very, very few of the derivations are explanatory. But this, I think, is no problem, for explanation is a false goal. (Cartwright, 'The Born-Einstein Debate: Where Application and Confirmation Separate' [29, page 281]).

Instead the goal –if any– is, in her view, *descriptive completeness.* I agree, but with one important qualification: we should not take the descriptive completeness of a theory always to argue for its empirical adequacy, or its truth. In my case study in superconductivity, in contrast to Cartwright's examples in [29], the theory helps to provide a description of a phenomenon that it simply cannot account for. The structure of application differs also dramatically from the structure of confirmation.

Let me briefly recapitulate. The case study in superconductivity points to the existence of a variety of ways in which scientific theories are applied. Some scientific applications are, perhaps, deidealizations of theory; some are reductions of theory in the sense of Friedman; and some are embeddable in theory *à la* Van Fraassen. I argue that there is a further class of applications that are not deidealizations of theory, are not reducible to theory, and cannot be embedded in theory. It is in those cases that mediating models play a key dual role. First, they help in the application of theory, by guiding the

introduction of corrections into the theory required in order to accurately describe the phenomena. Second, they provide us with physical insight into the nature of the phenomena. Because of that, mediating models are not just useful *fictions*; on the contrary they are carriers of very significant and specific knowledge of the phenomena. However, the role of mediating models in theory-application means that a realist construal of scientific theory becomes highly problematic. In these cases the theory itself is used as an instrument in application only and no attempt is made to confirm or disconfirm it at all.

In the first chapter I also sketched a distinction between *degree of confidence* and *degree of confirmation*. The London model does not raise the degree of confirmation of electromagnetic theory; it raises its *degree of confidence* –that is, it gives us a reason to believe that the theory is instrumentally reliable, i.e. that it will go on to provide successful applications. The instrumental reliability of a theory provides neither grounds to believe that the theory is true, nor that it is empirically adequate –it points neither towards scientific realism, nor towards constructive empiricism. Scientific realism and constructive empiricism share a common core, which is rejected by instrumental reliability. On either view a minimum requirement for the acceptance of a scientific theory is that the theory must be empirically adequate –i.e. that what the theory states is the case about the phenomena must indeed be the case. The constructive empiricist argues that the acceptance of a theory need only involve the belief that it is empirically adequate. Theories may have other virtues besides empirical adequacy –such as simplicity, explanatory power, aesthetic value, or even the virtue of being true... –, but belief in a theory's empirical adequacy is the only doxastic attitude required for the acceptance of the theory. By contrast, the realist argues that the belief that the theory is true, or likely to be true, and not just empirically adequate, is also required for its acceptance. For the realist a good theory, in addition to being empirically adequate, should also be true, or likely to be true –not only true to the phenomena, but true *tout court*, true to the world.

Thus, the scientific realist and her opponent, the constructive empiricist, agree that only highly confirmed theories should be accepted; we should have confidence in theories that are highly confirmed, and only in those. This is because on either view, confirmation always goes *via* empirical adequacy. A theory is confirmed when its observable predictions are borne out. The theory is empirically adequate if *all* the predictions of the theory, –past, present and future–, are borne out[17]. The realist takes a high degree of confirmation as a strong indication that the theory is true, or very likely to be true, because on her view empirical adequacy is a guide to truth. So, for the realist a high degree of confirmation is required for acceptance. For the constructive empiricist a high degree of confirmation is only an indication that the theory is empirically adequate, nothing more. But, as the constructive empiricist thinks that a belief in the empirical adequacy of a theory is required for its acceptance, he will readily agree with the realist that a high degree of confirmation is a requirement for accepting a theory.

According to Van Fraassen '*to accept a theory rather than another one involves also a commitment to a research programme, to continuing the dialogue with nature in the framework of one conceptual scheme rather than another*' (Van Fraassen, *The Scientific Image* [133, page 4]). And yet, the scientific community's commitment to classical electromagnetism as a programme of research seemed to require neither the belief that the electromagnetic theory was true, nor the belief that it was empirically adequate. Only a strong sense of confidence that the theory could be successfully applied was involved. Thus French and Ladyman can write, following Dahl [40]: "*the scientific community displayed a 'dogmatic faith' in the applicability to superconductivity of both thermodynamics and standard Maxwellian electromagnetic theory*" (French and Ladyman [68, page 26]).

To conclude: I have argued that, sometimes, in the decision to commit

[17]We may never be in a position to know if a theory is empirically adequate or not. The claim that a theory is empirically adequate carries *precisely the same commitment* to the correctness of a theory's future predictions, as does the claim that the theory is true. In this respect the constructive empiricist sticks his neck out *exactly as much as* the realist.

ourselves to a theory as a research programme, pragmatic considerations regarding its instrumental reliability may overrule epistemic considerations regarding its empirical adequacy. Sometimes, we have to settle for less[18].

2.6 Summary

In the previous chapter I introduced the notion of *instrumental reliability*. In this chapter I have focused on empirical adequacy, and I have explained the distinction between empirical adequacy and instrumental reliability. What makes a scientific theory empirically adequate? The answer crucially depends upon one's conception of what a scientific theory is. I have described two familiar views of the nature of scientific theories (section 2), and I have focused on the semantic conception of scientific theories, according to which theories are sets of models. I have reviewed the two most influential characterisations of empirical adequacy within the semantic view (section 3): a theory is empirically adequate if it embeds the phenomena (Van Fraassen), or if it reduces it (Friedman). I then have gone on to discuss the distinction, due to James Bogen and James Woodward, between data and phenomena; raised some difficulties for the semantic characterisations of empirical adequacy; and argued that this difficulties can be overcome by employing the notion of a *model of phenomena* (section 4). Finally (section 5) I have re-

[18] I say *sometimes* because I want to avoid, at all costs, the suggestion that instrumental reliability, rather than truth or empirical adequacy, is the *goal* of science. Far from it. In this respect I prefer to subscribe to Arthur Fine's Natural Ontological Attitude (NOA):

> Does science aim at truth, or does science merely aim at empirical adequacy? This is the springboard for the realism / instrumentalism controversy. NOA wants to pull back a bit from the question to ask, more fundamentally, whether science 'aims' at all. Of course, there is a point to particular investigations, and certainly particular research groups have aims and goals (to do a better experiment, to solve an outstanding problem, to build a better instrument, etc.). But only a fallacy in quantifier logic would lead one from 'They all have aims' to 'There is an aim that they all have'. (Fine [63, page 173]).

turned to the issue of the application of theories, and argued, by means of the case study in chapter 1, that a theory's domain of application is typically larger than its domain of empirical adequacy. I have concluded with some remarks on the significance of this fact for the scientific realism / antirealism debate.

Part II

Application in the Foundations of Quantum Theory

Chapter 3

Quantum Theory of Measurement

3.1 The Problem of Measurement

My aim in this chapter is to focus on a main issue in philosophy of physics where the distinction between application and confirmation matters. I shall argue that there are two problems of measurement in non-relativistic quantum mechanics. One has to do with the empirical adequacy of the quantum theory, the other with its application. I shall then describe Arthur Fine's solution to the problem of application, which, I shall argue, is not a solution to the problem of empirical adequacy. This should not come as a surprise in view of the discussion in Part I of the Thesis: the structure of application differs from that of confirmation, and we should not expect an application of a theory to always increase its degree of confirmation, or to display its empirical adequacy. Hence, I both offer a defence of Fine's solution to the measurement problem, and provide a diagnosis of the source of possible resistance to it. Fine proposes a novel account of the application of quantum theory to measurement interactions. But his proposal should not be taken as an attempt to solve the first measurement problem; and indeed from the point of view of empirical adequacy, and of confirmation, the proposal would

seem to be defective.

In a nutshell, the problem of measurement in non-relativistic quantum mechanics is the fact that the linear dynamics entailed by the Schrödinger equation is inconsistent with the requirement that measurements must have outcomes (i.e. that they must have *some* particular outcome or other). In this section I show that the measurement problem threatens to make the quantum theory empirically inadequate, given a standard interpretational rule, the *eigenstate-eigenvalue link.*

The essential difficulty raised by the problem of measurement can already be appreciated in the simpler formalism of quantum mechanics in terms of vectors in Hilbert space[1]. I explain the essential difficulty in section 3.1.1. However, a rigourous proof of the insolubility of the measurement problem can only be derived in the framework of statistical operators. The most general form of the problem of measurement is encapsulated in the so-called *insolubility proof* of the quantum measurement, which I derive in section 3.1.4. Section 3.1.2 concerns the ignorance interpretation of mixtures, which is essential for the whole chapter, while section 3.1.3 is preparatory for the discussion of the insolubility proof.

3.1.1 The Measurement Problem for Pure States

Consider an object system O initially in state $|\phi\rangle \in H_1$. We are interested in the values of some observable O of this system, represented by the Hermitian operator[2] $\hat{O}$. We make measurements on this system by letting it interact with a measuring apparatus M initially in state $|R_0\rangle \in H_2$. We then measure the 'pointer position' observable, represented by $\hat{A}$, on the measuring apparatus M. We can represent the initial combined state of the

[1]See Appendix 1, section 3.5. Throughout this chapter I shall make much use of some established technical results in the formalism of quantum theory, which I review in the Appendixes.

[2]See Appendix 1. In this Thesis I use the terms 'observable' and 'operator' exchangeably. Which one is which should be clear from the context.

system and the apparatus in a larger Hilbert space $H_{1+2} = H_1 \otimes H_2$, formed as the tensor product of the smaller spaces[3]. Suppose that the eigenstates of $\hat{O}$ are the set $\{\phi_i\}$, with corresponding eigenvalues λ_i, while the eigenstates of $\hat{A}$ are $\{R_i\}$ with eigenvalues a_i, i.e.:

$$\hat{O}|\phi_1\rangle = \lambda_1|\phi_1\rangle, \hat{O}|\phi_2\rangle = \lambda_2|\phi_2\rangle, etc.$$
$$\hat{A}|R_1\rangle = a_1|R_1\rangle, \hat{A}|R_2\rangle = a_2|R_2\rangle, etc.$$

Suppose further that the initial state of the object system is one of the eigenstates $|\phi_i\rangle$ of the operator $\hat{O}$. The initial state of the composite system is then given by the vector $|\phi_i\rangle \otimes |R_0\rangle$ in the tensor product space. We are interested in the value that observable $(O \otimes A)$, defined on $H_1 \otimes H_2$, takes on the final evolved state of the composite system.

We now impose the following minimal condition on any satisfactory measurement interaction U_t, namely that it should correlate the initial states of the object system with distinguishable states of the apparatus system (we don't require that the interaction leave the object system state unchanged):

$$|\phi_i\rangle \otimes |R_O\rangle \rightarrow |\phi_i'\rangle \otimes |R_i\rangle.$$

It is then possible to show that if the initial state of the object system is a superposition over some or all of the eigenstates of $\hat{O}$ then the final evolved state of the composite system at the end of the measurement interaction, is a superposition over eigenstates of $(\hat{O}' \otimes \hat{A})$. If $|\phi\rangle = \sum_i c_i|\phi_i\rangle$ then:

$$|\phi\rangle \otimes |R_O\rangle = \sum_i c_i|\phi_i\rangle \otimes |R_O\rangle \rightarrow \sum_i c_i|\phi_i'\rangle \otimes |R_i\rangle. \qquad (3.1.1)$$

The problem of measurement can now be generated by attending to a standard, interpretational principle of quantum mechanics, the so-called *eigenstate-eigenvalue link*, or *e/e link* for short[4]. According to the *e/e link*

[3]See Appendix 2 for a discussion of the interaction formalism of Quantum Theory.

[4]Although implicit throughout the history of the foundations of quantum mechanics, the eigenstate-eigenvalue link was first explicitly introduced by Arthur Fine [58] only as late as 1973. Fine also showed [59] that the *e/e link* cannot be *derived*, as Von Neumann thought, from further principles of the quantum theory, but that it needs to be assumed.

an observable of a quantum system can be said to have a value *if and only if* the system is in an eigenstate of the Hermitian operator that represents the observable. It follows that the observable $(I \otimes A)$ has no value in the state of the composite object-apparatus system that results at the end of the 'measurement' interaction (3.1.1). The final evolved state of the composite is not an eigenstate of the required observable $(I \otimes A)$ but, rather, a superposition of such eigenstates. Quantum theory predicts, then, that measurement interactions will normally fail to have any outcomes. If, as it is the case, the occurrence of outcomes of quantum measurement interactions is a well-established phenomenological fact, quantum theory turns out to be empirically inadequate.

3.1.2 The Ignorance Interpretation of Mixtures

In the most general formalism of quantum theory, states are represented by Hermitian operators acting on the Hilbert space[5]. This formalism applies to *mixed* as well as pure states. A mixed state[6] is a state of '*less than maximal information*'. For any pure state $|\phi\rangle$ there is in principle an observable that can test whether a system is in this state, namely the observable represented by the projector $P_{[|\phi\rangle]}$. A measurement of this observable on any system can have one of two outcomes, 0 or 1; only if the system is in state $|\phi\rangle$ will the outcome be 1 with probability 1. There is no similar 'testing' observable for mixed states.

We represent a mixed state W as a virtual ensemble of pure states $\{|\phi_i\rangle\}$ with statistical weights $\{p_i\}$, $0 \leq p_i \leq 1$, where $\sum_i p_i = 1$:

$$\begin{aligned} W : p_1 &: |\phi_1\rangle \\ p_2 &: |\phi_2\rangle \\ &\cdots \\ p_n &: |\phi_n\rangle \end{aligned}$$

[5]See Appendix 2.

[6]See Fano [54, page 74].

More generally we represent mixed states by so-called statistical operators: $\hat{W} = \sum_i p_i |\, v_i \rangle\langle\, v_i\,| = \sum_i p_i \hat{P}_{|v_i\rangle}$, where the set of pure states $\{|v_i\rangle\}$ can be any set you like, and *need not* be orthogonal or complete. However, statistical operators are Hermitian and the Spectral Decomposition Theorem[7] dictates that it must be possible to write $\hat{W}$ in terms of mutually orthogonal projectors, regardless of its initial description.

I have defined a quantum mixed state as a *virtual* ensemble of pure states $\hat{P}_{[v_i]} = |\, v_i \rangle\langle\, v_i\,|$, with associated statistical weights p_i. A mixed state in classical statistical physics represents a possible state of a real ensemble of systems –each member of the ensemble being in a particular (pure) state. In such case, the statistical weights represent the relative frequencies in the overall ensemble of systems in each state. Hence, in classical physics mixed states cannot be ascribed to individual systems, but only to actual collections of systems. By contrast, according to many authors, quantum mechanical mixed states can be ascribed to individual systems as well as collections. Thus quantum mechanical mixed states can correspond to *virtual*, as well as *real* ensembles of systems.

In 1948, Hans Reichenbach [116] proposed the *ignorance interpretation* of quantum mixtures. Reichenbach's aim was to construe quantum mixtures as an expression of our incomplete knowledge of the actual state of the system. He suggested that the statistical weights p_i represented not an intrinsic property of the system, but rather our ignorance of the state of the system in question. According to Reichenbach's definition a system is in a state represented by $\hat{W} = \sum_i p_i |\, v_i \rangle\langle\, v_i\,|$ if and only if the system is *really* in one of the pure states $|v_i\rangle$, the weights p_i representing our degree of ignorance.

There are two standard arguments against the ignorance interpretation, and one in its favour. In favour of the ignorance interpretation is the fact that it is consistent with the dynamical evolution of mixed states. If the initial mixture is $W = \sum_i p_i |\, v_i \rangle\langle\, v_i\,|$, and the final, time-evolved, mixture is $W = \sum_i p_i |\, u_i \rangle\langle\, u_i\,|$ we expect, given the ignorance interpretation, that

[7]Theorem (3.5.1) in Appendix 1.

there must be a unitary operator $\hat{U}_t$ such that:

$$\forall i : |u_i\rangle = \hat{U}_t|v_i\rangle$$

Hence:

$$W(t) = \sum_i p_i|\, u_i\rangle\langle\, u_i\,| = \sum_i p_i\hat{U}_t|\, v_i\rangle\langle\, v_i\,|\hat{U}_t^{-1} =$$

$$\hat{U}_t\left(\sum_i p_i|\, v_i\rangle\langle\, v_i\,|\right)\hat{U_t^{-1}} = \hat{U}_t\hat{W}\hat{U_t^{-1}}$$

which is indeed the time-evolution of mixed states.

On the other hand, the non-uniqueness of a mixed state's decomposition has been thought to militate against the ignorance interpretation. If a mixture takes the form $\hat{W} = \sum_i w_i\hat{P}_{|w_i\rangle}$, where the $|w_i\rangle$ are not necessarily orthogonal, and/or $w_i = w_j$ for some i, j, then there will exist at least two different decompositions, as a spectral decomposition always exists:

$$\hat{W} = \sum_i w_i\hat{P}_{|w_i\rangle} = \sum_i p_i\hat{P}_{|v_i\rangle},$$

where the $\{|v_i\rangle\}$ are mutually orthogonal.

The ignorance interpretation cannot be given to both decompositions simultaneously[8]. We need to choose one –but which one? The orthodox answer has been always to privilege the spectral decomposition in terms of orthogonal projectors. But as Paul Feyerabend [55] was perhaps the first to point out, in cases of degeneracy, where $\exists i, j : p_i = p_j$, even the orthogonal decomposition is not unique. Nancy Cartwright [25] finally refuted the orthodox answer, by displaying one real physical interaction, with the system and the measuring apparatus modelled in Hilbert spaces of different dimensions, where the orthodox answer leads to results out of keeping with the usual analysis of the interaction.

D'Espagnat[9] distinguished between *proper* and *improper* mixtures. $W = \sum_i w_i|\, w_i\rangle\langle\, w_i\,|$ is a proper mixture (a mixture 'by construction') if it is the

[8]See, for instance, Van Fraassen [138, page 206].

[9]D'Espagnat [44, section 7.2].

result of a preparation procedure that ensures us that the correct decomposition is over the set of $|w_i\rangle$, not $|v_i\rangle$. By contrast an improper mixture (a mixture 'by derivation') is ascribed to a subsystem of a composite system by deriving it from the state of the composite, and it carries no similar assurance as to the preferred decomposition. D'Espagnat concluded that proper mixtures can be said always to admit the ignorance interpretation, while improper mixtures cannot generally be said to admit that interpretation.

The decisive argument against the ignorance interpretation of *improper* mixtures is as follows (a similar, but not identical, argument appears in Hughes [89, pages 149-151]). Consider a composite system S_{1+2} in a pure state $W_{1+2} = |\Psi\rangle\langle\Psi|$, where $|\Psi\rangle = \sum_{i,j} c_{ij}|v_i\rangle \otimes |w_j\rangle$, and where, as in Appendix 3, $|v_i\rangle$ ($|w_j\rangle$) are the eigenstates of A (B) with corresponding eigenvalues a_i (b_j). This is clearly a pure state, as $Tr(W_{1+2}^2) = Tr(W_{1+2}) = \sum_{ij} |c_{ij}|^2 = 1$.

The *reduced states* W_1, W_2 can be derived from the standard identifications[10] given by (3.7.11):

$$Tr((\hat{A} \otimes \hat{I})\hat{W}_{1+2}) = Tr(\hat{A}\hat{W}_1)$$
$$Tr((\hat{I} \otimes \hat{B})\hat{W}_{1+2}) = Tr(\hat{B}\hat{W}_2)$$

We obtain: $W_1 = \sum_i c_{ii}c_{ii}^*|v_i\rangle\langle v_i|$, and $W_2 = \sum_j c_{jj}c_{jj}^*|w_j\rangle\langle w_j|$. Thus W_1 and W_2 are *improper* mixtures, found by derivation from the composite state W_{1+2}. Let us now assume that subsystem S_1 (S_2) is really in one of the states $|v_i\rangle\langle v_i|$ ($|w_j\rangle\langle w_j|$) with the probabilities $|c_{ii}|^2$ ($|c_{jj}|^2$). The state of the combined system can then be reconstructed, in the manner described in Appendix 3, section 3.7. We find that $W_{1+2} = \sum_i |c_{ii}|^2|v_i\rangle\langle v_i| \otimes W_2$ (or $W_{1+2} = \sum_j W_1 \otimes |c_{jj}|^2|w_j\rangle\langle w_j|$, or if *both* W_1 and W_2 are given the ignorance interpretation then: $W_{1+2} = \sum_{ij} |c_{ii}|^2|c_{jj}|^2|v_i \otimes v_j\rangle\langle v_i \otimes w_j|$).

Thus, on the assumption that W_1 (or W_2, or both) can be given the ignorance interpretation, we find that W_{1+2} is itself a mixture. But by

[10] See Appendix 3 for the details.

hypothesis W_{1+2} is a pure state; therefore, by *reductio ad absurdum*, the ignorance interpretation cannot in general be given to *improper* mixtures.

3.1.3 Conditions on Measurement Interactions

A model of quantum measurements will be expected to treat the interaction of the measured object with the measuring device. Quantum theory does indeed tell us how to construct an interaction model, in the tensor-product Hilbert space formalism. In order to measure a certain property of the object system we need to record a correlated property of the measuring device; the problem of measurement becomes the problem of the *objectification*[11] of the pointer position observable. A measurement of a particular property is usually understood as a mapping of objects, or states of objects, into the real numbers; a function M from a set of objects S onto the reals $\mathbf{R}$, $M : S \rightarrow \mathbf{R}$. Does the observable represented by $\hat{I} \otimes \hat{A}$ (the *pointer position observable*) lay out the appropriate kind of mapping into $\mathbf{R}$ characteristic of a measurement? Is it possible to ascribe a value to the position of the pointer of the measuring device at the end of the interaction? For an arbitrary initial state of the object system, an established family of results in the literature (the so-called *insolubility proofs* of the measurement problem by Wigner [140], Earman and Shimony [51], Fine [56] and [57], and Brown [18]) shows that a consistent mapping $M : S \rightarrow \mathbf{R}$ from all possible final states of the composite (*object + apparatus*) system into the real numbers is in general impossible.

Let us refer to the initial state of the object as W_o, and to the initial state of the apparatus W_a. In order to perform a measurement of the state of the quantum object by the measuring apparatus, an interaction must be set up between the object and the apparatus. This interaction will be governed by an appropriate Hamiltonian H. According to Stone's Theorem, there is a one-parameter group of unitary operators $\hat{U}_t$ associated with this Hamiltonian ($\hat{U}_t = e^{i\hat{H}t/\hbar}$). This unitary operator will then dictate the

[11] The term is due to Busch, Lahti and Mittelstaedt [21, page 75-83].

evolution of the composite state while the interaction takes place:

$$\hat{U}_t : W_o \otimes W_a \rightarrow \hat{U}_t(W_o \otimes W_a)\hat{U}_t^{-1}$$

We would like to establish the conditions that a unitary operator $\hat{U}_t$ must satisfy to qualify as a measurement interaction. Suppose that $(\hat{I} \otimes \hat{A})$, with eigenvalues μ_n and eigenvectors $|\beta_n\rangle$, represents the pointer position observable, and $\hat{O}$, with eigenvalues λ_i and eigenvectors $|\phi_i\rangle$ represents the observable of the object system ($\hat{O}$ and $\hat{I} \otimes \hat{A}$ are discrete, but may be degenerate). Consider the following two conditions on unitary operators, which I call the *Transfer of Probability* condition (TPC) and the *Occurrence of Outcomes* condition (OOC)[12]:

Definition 1 (Transfer of Probability Condition (TPC))

$$P(W_o, O) = P(\hat{U}(W \otimes W_a)\hat{U}^{-1}, I \otimes A)$$

(TPC) expresses the requirement that the probability distribution over the possible outcomes of the relevant observable O of the object system, be reproduced as the probability distribution over possible outcomes of the pointer position observable in the final state of the composite (*object + apparatus*) system. (TPC) differs from the *Probability Reproducibility* condition [21, page 32] of Busch, Lahti and Mittelstaedt[13], and is closer in spirit to the weaker equivalence condition (MEAS) adopted by Fine [57], and Brown [18] which I shall discuss below.

[12]These conditions involve the generalised Born Rule for statistical operators (3.6.10), which I describe in Appendix 2. In stating these conditions I have slightly varied the terminology used in the Appendix, in order to deal with statistical distributions over eigenvalues: $P(W, A)$ is the distribution defined by calculating, for each eigenvalue λ of $\hat{A}$, the probability $Prob_W(A = \lambda)$.

[13](TPC) and the Probability Reproducibility Condition are different, yet *essentially equivalent* conditions. Busch, Lahti and Mittelstaedt require that the probability distribution for the required observable in the initial state of the object system be reproduced in the probability distribution for the pointer observable in the final *reduced state* of the apparatus. Suppose that W_a^f represents the final *reduced* state of the apparatus, derived from the final composite state $\hat{U}(W_o \otimes Wa)\hat{U}^{-1}$ by the standard identifications (3.7.11).

Definition 2 (Occurrence of Outcomes Condition (OOC))

$$U_t(W_o \otimes W_a)U_t^{-1} = \sum_n c_n W_n$$

where for every W_n *there is some eigenvalue* μ_n *of* $\hat{I} \otimes \hat{A}$ *such that* $P(W_n, I \otimes A = \mu_n) = 1$

The thought that underlies (OOC) is that we must require, in addition, that our models of measurement account for the phenomenological fact that measurement interactions typically have outcomes. Quantum theory applies to individual particles as well as to collections of particles. We need to require not only that the probability distributions be transferred but also that, when a single measurement is made on one system, an outcome be found. While (TPC) ensures that the statistical distribution for the relevant observable O will be faithfully revealed by measurements on an ensemble of measuring devices, it does not guarantee that single measurements on individual particles will have outcomes. (OOC) guarantees that on measuring some quantity of an individual particle, an outcome (some outcome, not one particular outcome) will be displayed on the measuring apparatus.

(OOC) is necessary for an ignorance interpretation of the final state of the composite. The ignorance interpretation allows us to say that when a system ends up in a mixture like the one described in (OOC), the system is really in one of the pure states that make up the mixture, but *we don't know which one.* We would naturally write down its state in the form $\sum_n c_n W_n$, (where W_n are the possible states of the system) if we had incomplete knowledge of the system's state; the numbers c_n $(0 \leq c_n \leq 1)$ would then represent the probabilities that the system really is in each of W_n. As each of W_n is a state for which $(I \otimes A)$ takes some value with probability

The Probability Reproducibility Condition reads:

$$Prob(W_o, O) = Prob(W_a^f, A)$$

which, given the derivation of the reduced state W_a^f from the final state of the composite by means of (3.7.11), is provably equivalent to (TPC) for observable A.

one, we can be confident that a system in state $\sum_n c_n W_n$ will always display a definite outcome, *although we don't know which one.*

There is an important assumption buried in this last remark. I have assumed that if $(I \otimes A)$ is to have a value in state W_n, the probability of finding the outcome μ_n on a measurement of $(I \otimes A)$ is one –where μ_n is an eigenvalue of the operator $\hat{I} \otimes \hat{A}$. In other words, I have assumed that an extension of the *e/e link* holds for mixed states. Arthur Fine[14] has provided one such extension, which is captured in his two interpretational rules for the ascription of values to observables in mixed states, the *Rule of Law* and the *Rule of Silence.* The *Rule of Law* is the equivalent of the sufficient part of the *e/elink*, while the *Rule of Silence* is equivalent to its necessary part. Together the *Rule of Law* and the *Rule of Silence* amount to a complete generalisation of the *e/elink* to mixed states, which I state as follows[15]:

Definition 3 (Generalised Eigenstate/Eigenvalue Link) *(a) Observable Q has a value in state W_n if and only if there is an eigenvalue λ_n of Q, such that $P(W_n, Q)(\lambda_n) = 1$, in which case Q takes the value λ_n. (b) If the state W of a system is a mixture over states W_n, Q has a value in W if and only if Q has a value in each of W_n in the sense of part (a).*

I shall assume that (TPC) and (OOC) are jointly sufficient for an interaction to count as a measurement. Indeed (OOC) is, on its own, a necessary condition on measurement interactions. For as long as the *generalised e/e link* is in place, and as long as we allow that the final state of the composite may be a mixed state, we are not entitled to say that measurement interactions have outcomes, unless (OOC) holds. In contrast, (TPC) is not a necessary condition. A weaker requirement on the transfer of probabilities (one, for instance, requiring less than the transfer of the *entire* probability distribution) may be enough for the interaction to reveal some information

[14] Fine [64], [66].

[15] Note that part (b) of this generalised e/e link involves the ignorance interpretation of mixtures (Fine [64, page 496]), as required by the Occurrence of Outcomes Condition.

regarding the initial state of the object system. Thus, although it is true that any unitary operator that satisfies the conjunction (TPC) & (OOC) qualifies as a measurement interaction, we may well be leaving out some further interesting, class of measurement interactions. Indeed Fine [57, page 2784] has given an equivalence condition that, he claims, is the minimal necessary condition for a measurement (exactly the same condition is also employed by Brown [18]). Consider the following definition of state equivalence, relative to an observable Q:

Definition 4 (Q-Equivalence) *Two states W and W' are Q-equivalent if and only if $prob(W,Q) = prob(W',Q)$, in which case we write: $W \equiv_Q W'$*

This definition generates equivalence classes of states (pure or mixed) with respect to some observable; it groups all those states with identical probability distributions over the eigenvalues of a particular observable. As a matter of fact, the definition ensures that the states W and W' are statistically indistinguishable and thus, on the standard understanding of quantum mechanics, empirically indistinguishable with respect to the observable Q. We may thus refer to states W, W' for which the above Q-equivalence holds, as Q-indistinguishable states. We can now write down Fine's minimal necessary condition on measurement interactions. Consider a measuring apparatus in the initial state W_a, and two possible initial states of the object system, W_o and W'_o. Then:

Definition 5 (Measurement (MEAS)) *A Unitary Operator $\hat{U}$ is a W_a measurement (of the observable O by means of the observable A) if and only if whenever $\hat{U}(W_o \otimes W_a)\hat{U}^{-1} \equiv_{I\otimes A} \hat{U}(W'_o \otimes W_a)\hat{U}^{-1}$ then $W_o \equiv_O W'_o$.*

A *measurement* is an interaction that yields $I \otimes A$-distinguishable final states of the composite whenever the initial states of the object system are O-distinguishable. It is not required by this condition that the unitary operator be capable of transferring the entire probability distribution from the object system to the composite, but only that if two initial states of the object system are statistically distinguishable with respect to O, their corresponding

time-evolved composite states, for any arbitrary initial state of the apparatus, be distinguishable with respect to $I \otimes A$. The pointer position observable must be able to make out these two final states of the composite; but it is not required that the distribution over O fixed by the initial object-state, be reproduced as the distribution over $I \otimes A$ fixed by the corresponding evolved state of the composite. So Fine's condition on measurements (MEAS) does not entail (TPC). It is clear, though, that (TPC) entails Fine's minimal condition. (TPC) requires the entire probability distribution over eigenvalues of $\hat{O}$ in the initial state of the system to be reproduced as the distribution over eigenvalues of $\hat{I} \otimes \hat{A}$ in the final state of the composite. Hence, given a fixed state W_a of the apparatus, any difference in the probability ascribed to one of the eigenvalues of $\hat{I} \otimes \hat{A}$ by two different final states of the composite necessarily reflects some difference between probabilities ascribed to different eigenvalues of $\hat{O}$ by the two corresponding initial states of the object.

3.1.4 The Insolubility Proof

The essence of the measurement problem is constituted by a proof that shows (OOC) and the minimal (MEAS) condition on measurements to be inconsistent for some sets of initial states of the object system. As (TPC) entails (MEAS), the same proof automatically shows that (TPC) and (OOC) are inconsistent, and for precisely the same sets of initial states of the object system. I focus on the most general version of the proof, originally due to Arthur Fine [57], and slightly reformulated by Harvey Brown [18].

For the sake of generality, let the initial state of the measuring apparatus be $W_a = \sum w_n \hat{P}_{[\gamma_n]}$, i.e. a mixture over pure states γ_n. Consider three O-distinguishable initial states of the object system, namely $\hat{P}_{[\phi_1]}$, $\hat{P}_{[\phi_2]}$ and $\hat{P}_{[\phi_3]}$, where ϕ_1 and ϕ_2 are eigenstates of $\hat{O}$, with eigenvalues λ_1 and λ_2, but ϕ_3 is a linear combination of both: $\phi_3 = a_1\phi_1 + a_2\phi_2$. It is clear that ϕ_1, ϕ_2 and ϕ_3 are O-distinguishable. Let us now set up an interaction in accordance with the Schrödinger equation, represented by a unitary operator

$\hat{U}$, for which (OOC) is satisfied. Hence we have for ϕ_1, ϕ_2, ϕ_3:

$$\hat{U}\left(\hat{P}_{[\phi_1]} \otimes W_a\right)\hat{U}^{-1} = \sum w_n \hat{P}_{[\beta_{n1}]}$$
$$\hat{U}\left(\hat{P}_{[\phi_2]} \otimes W_a\right)\hat{U}^{-1} = \sum w_n \hat{P}_{[\beta_{n2}]}$$
$$\hat{U}\left(\hat{P}_{[\phi_3]} \otimes W_a\right)\hat{U}^{-1} = \sum w_n \hat{P}_{[\beta_{n3}]}$$

where $\beta_{ni} = \hat{U}(\phi_i \otimes \gamma_n)$.

In order to derive this equality, it is necessary to appeal to a law of evolution named *Real Unitary Evolution* (RUE) by Harvey Brown [18, page 860]. (RUE) dictates that a mixture $W = \sum_n w_n \hat{P}_\phi$ is evolved by unitary evolution into the mixture $W^f = \hat{U}W\hat{U}^{-1} = \sum_n w_n \hat{U}\hat{P}_{[\phi]}\hat{U}^{-1} = \sum_n w_n \hat{P}_{\hat{U}\phi}$. Hence (RUE) asserts that the evolution of a mixture over pure states is the mixture over the evolved pure states. It is easy to see that (RUE) is consistent with the positive argument in favour of the ignorance interpretation given in 3.1.2. In fact, (RUE) follows naturally from the ignorance interpretation of mixtures. However, (RUE) does not require the ignorance interpretation. Although it is arguably harder to find a plausible rationale for (RUE) in the absence of such interpretation, it is nevertheless possible to postulate it independently.

Now we have, by the linearity of U, that:

$$\hat{U}(\phi_3 \otimes \gamma_n) = a_1\hat{U}(\phi_1 \otimes \gamma_n) + a_2\hat{U}(\phi_2 \otimes \gamma_n)$$

or:

$$\beta_{n3} = a_1\beta_{n1} + a_2\beta_{n2}.$$

If (OOC) is to hold it must be the case that:

$$\forall n \forall i = 1,2,3 : \hat{I} \otimes \hat{A}(\beta_{ni}) = \mu_{ni}\beta_{ni},$$

where μ_{ni} are eigenvalues of $\hat{I} \otimes \hat{A}$. But now we can see that (OOC) and the minimal condition on measurements (MEAS) cannot simultaneously hold. For:

$$\hat{I} \otimes \hat{A}(\beta_{n3}) =$$

$$\hat{I} \otimes \hat{A}(a_1\beta_{n1} + a_2\beta_{n2}) =$$
$$a_1(\hat{I} \otimes \hat{A})\beta_{n1} + a_2(\hat{I} \otimes \hat{A})\beta_{n2} =$$
$$a_1\mu_{n1}\beta_{n1} + a_2\mu_{n2}\beta_{n2}, \tag{3.1.2}$$

and also:

$$\hat{I} \otimes \hat{A}(\beta_{n3}) = \mu_{n3}\beta_{n3} =$$
$$\mu_{n3}(a_1\beta_{n1} + a_2\beta_{n2}) =$$
$$a_1\mu_{n3}\beta_{n1} + a_2\mu_{n3}\beta_{n2}. \tag{3.1.3}$$

However, (3.1.2) and (3.1.3) can only hold if $\mu_{n1} = \mu_{n2} = \mu_{n3}$. Thus, the final states of the composite $\sum w_n \hat{P}_{[\beta_{ni}]}$ are not $I \otimes A$-distinguishable, contradicting (MEAS), the minimal necessary condition on measurements, and hence contradicting (TPC). It is therefore impossible, under the assumption of the *generalised e/e link*, to set up an interaction between an object system and a measuring apparatus that simultaneously obeys the Schrödinger equation, the Transfer of Probability Condition (TPC), and the Occurrence of Outcomes Condition (OOC): the measurement process cannot be modelled entirely within quantum theory, given its standard interpretation.

3.2 The Modal Interpretation and Its Problems

The insolubility proof poses a conundrum. By assuming (TPC), (OOC), the Schrödinger dynamics, and the *generalised e/e link* we can generate a measurement problem. Hence, at least one among these assumptions must be false, if the measurement problem is to have a solution[16]. (OOC) and (TPC) are constitutive of measurement interactions, so it seems that in order to escape the insolubility result we are required to either (1) change, or

[16]In section 3.3 I shall argue, following Arthur Fine, that a further assumption is involved in the insolubility proof. This fifth assumption is related to the *application* of the quantum interaction formalism. For the time being, however, and for the sake of argument, assume that the only assumptions involved are the Schrödinger dynamics, the *generalised e/e link*, and the conditions (TPC) and (OOC).

supplement, the Schrödinger dynamics, or (2) relinquish the standard interpretative principle, the *generalised e/e link*. The former way out (advocated originally by Von Neumann [109], and more recently by Ghirardi, Rimini and Weber [72], and in a different way by Gisin [75]) involves changing the quantum theory altogether, as it implies a revision of one of its basic principles. In the end, all factors considered, this drastic solution may well be the best response to the measurement problem. However, for the purposes of this Thesis I am interested in philosophical attempts to assess the empirical adequacy of the theory and to describe its domain of application. Hence, in this section I shall only be concerned with attempts to provide a solution to the measurement problem by relinquishing the *e/e link*.

These are, in short, attempts to show that the threat against the empirical adequacy of the quantum theory is only apparent. This appearance is the by-product of a standard, but ultimately mistaken, interpretation of the theory; when the correct interpretation is given, the threat will simply disappear. There are a number of proposals of this kind in the literature, most of which spring from Everett's work [53]. I shall address only what I consider the best developed attempt in this direction: the modal interpretation of quantum theory.

3.2.1 The Kochen-Healey-Dieks Modal Interpretation

The modal interpretation of quantum theory originates in work by Bas Van Fraassen in the early 1970's (see [130], [131]). Van Fraassen's intuition was (and still is) that the measurement problem can be resolved by introducing a new *interpretational* rule for the ascription of values to observables. The modal interpretation effectively relinquishes the necessary part of the *e/e link*. It then becomes possible for an observable to have a value even when the state of the system is not an eigenstate of the Hermitian operator that represents the observable. We can then dispense with (OOC) as a necessary condition altogether. It becomes unnecessary, for measurements to have outcomes, that the final state of the (*object + apparatus*) composite system be a mixture over eigenstates of the pointer position observable. The pointer

will have a value, and the measurement will have an outcome, even if the the composite ends up in a superposition.

A *state* fulfils two roles in a physical theory. First, it determines what the properties of the system are, and what values those properties have. Second, it determines how the values of those properties will evolve, if the system is isolated. A classical mechanical state will fulfil these two functions simultaneously. In the modal interpretation of quantum mechanics, however, systems are said to possess two states –which Van Fraasen[17] calls the *dynamic state* and the *value state*– and the two functions are split. Modal interpretations of interactive systems ascribe to the composite system *both* a dynamic state, represented by a statistical operator W_{1+2} acting on the tensor product Hilbert space $H_1 \otimes H_2$, and a value state represented by a vector in that space. The dynamic state is fully specified by stating how the system will develop. This is the 'usual' quantum-mechanical state: its evolution is governed by the Schrödinger equation, and can thus be represented by the action of a unitary operator $\hat{U}_{1+2}$ on the tensor product Hilbert space. On the other hand, the value state has no Schrödinger evolution. At any particular time, the value state is fully specified by stating which observables have values and what those values are. Hence, the value state of a composite system can be represented by a vector in the tensor product Hilbert space.

The subsystems of a composite can be ascribed their own dynamic and value states. Their dynamic states will be represented by means of statistical operators W_1 and W_2 acting on the individual Hilbert spaces H_1 and H_2, obtainable from the overall state W_{1+2} of the composite by applying the standard identifications (3.7.11). W_1 and W_2 will typically be non-idempotent operators, i.e. they will represent mixtures. We expect their evolution to be given by unitary operators $\hat{U}_1$ and $\hat{U}_2$, acting on the spaces H_1 and H_2 (but on this issue, more to follow). On the other hand, each subsystem of the composite can also, at any time, be ascribed an individual

[17] Van Fraassen [138, page 275].

value state. This will be a pure state, represented by a vector in the corresponding Hilbert space H_1 or H_2. The set of all possible value states of a subsystem is determined by the spectral resolution of its individual dynamic state, W_1 or W_2.

If the state of the composite is pure then the value states of the subsystems figure in the so-called biorthonormal decomposition of the combined state:

$$|\psi\rangle = \sum_i c_i |\psi_i\rangle \otimes |\eta_i\rangle,$$

where $|\psi_i\rangle \in H_1$, and $|\eta_i\rangle \in H_2$. The case for the biorthonormal decomposition has been argued most forcefully by Simon Kochen, Richard Healey and Dennis Dieks[18]. The biorthonormal decomposition theorem guarantees that every state in the tensor product space always has one such decomposition; this decomposition is moreover unique, except in the case of degeneracies[19]. So, at any time during the interaction the subsystems may be ascribed reduced mixed states, represented by density operators W_1 on H_1 and W_2 on H_2, whose spectral decompositions range precisely over their respective subsystem's possible value states. These facts about the modal interpretation can be summarised in the following diagram:

$$\begin{array}{ccc}
|\psi\rangle = \sum_i c_i|\psi_i\rangle \otimes |\eta_i\rangle & \xrightarrow{\hat{U}_{1+2}} & |\psi^t\rangle = \sum_i d_i|\phi_i\rangle \otimes |\xi_i\rangle \\
\Big\downarrow \textit{decomp} & & \Big\downarrow \textit{decomp} \\
W_1 = \sum_i |c_i|^2 |\psi_i\rangle\langle\psi_i| & & W_1^t = \sum_i |d_i|^2 |\phi_i\rangle\langle\phi_i| \\
\vdots & & \vdots \\
W_2 = \sum_i |c_i|^2 |\eta_i\rangle\langle\eta_i| & & W_2^t = \sum_i |d_i|^2 |\xi_i\rangle\langle\xi_i|
\end{array}$$

[18] Kochen [95], Dieks [45] and [46] and Healey [82].

[19] See Kochen [95], page 154.

As the biorthonormal decomposition is unique, it effectively selects a privileged form of representation for W_1 and W_2. In section 3.1.2 I mentioned that a well-known argument against the ignorance interpretation is precisely that mixtures generally have no unique representation. For an arbitrary mixed state $W = \sum_i w_i|\psi_i\rangle\langle\psi_i|$ which might be degenerate and where the set $\{|\psi_i\rangle\}$ is not necessarily pairwise orthogonal, there are at least two different representations, as one is always guaranteed by the spectral decomposition theorem. If the mixture is improper then there is no way to tell which representation is physically correct, so we cannot apply ignorance. But, on the other hand, the lack of uniqueness would cease to be a problem if we had an independent rule to determine the physically correct representation. The biorthonormal decomposition is one such rule. It justifies the ignorance interpretation of the reduced states W_1, W_2 of the subsystems, because it makes the representation of those states unique[20]. It then becomes permissible to give W_1 and W_2 an ignorance interpretation. That is, we may assume that the subsystems *really* possess one of the pure states in the mixture with the associated probabilities given by the square norms

[20]This statement needs to be heavily qualified in light of the discussion in section 3.1.2. For a start, the biorthonormal rule will always privilege the orthogonal decompositions for the reduced states. In section 3.1.2 I referred to this as the orthodox choice for the representation of mixtures, which Paul Feyerabend and Nancy Cartwright have shown won't work as a *universal* rule for applying ignorance. Hence, the biorthonormal decomposition rule justifies the ignorance interpretation as long as we assume that (1) the dimensions of the Hilbert spaces of the subsystems is the same, and (2) there are no degeneracies in the reduced states. The latter is possible only if the pure state of the composite is totally non-degenerate. Dieks initially argued that degenerate states of the composite form a set of measure-zero in the total set of possible states, and hence can be ignored [45, page 443]. As will become clear soon, I think that measure-zero arguments, whether in favour or against modal interpretations, are generally unsound, unless the particular measure in question can be given some direct physical motivation. But in this instance such motivation is absent, as is emphasised by the fact that the most discussed ever composite state in philosophy of quantum mechanics, the EPR singlet state, is totally degenerate. Dieks now favours a *decoherence* solution to this problem: see, for instance, his [47]. Bacciagaluppi and Hemmo [7] have further developed the decoherence solution.

of the coefficients. I am not suggesting that the ignorance interpretation should be given –in fact I shall shortly argue to the contrary. But the suggestion is rather obvious: in the modal interpretation of interactive systems it is always possible to represent the set of value states of the subsystems by means of the pure states that appear in the spectral decomposition of those mixtures. Hence the modal interpretation of the reduced mixtures is just like the ignorance interpretation. As Van Fraassen writes:

> *It is as if the ignorance interpretation of mixtures were correct.* For if system X is in mixed state W, then the actual values of observables pertaining to X are exactly those it would have had if it had been a pure state in [the spectral resolution] of W. But we don't know which pure state –all of them are possibilities for us, if we are told only that X is in mixed state W. (Van Fraassen, [138, page 22])

And yet, as we can see, Van Fraassen actually refrains from fully embrading the ignorance interpretation of mixtures. In the course of this discussion, I shall argue that this is a wise thing to do: it is indeed impossible, in general, to give an ignorance interpretation to the reduced mixtures W_1 and W_2 that represent the dynamic states of the subsystems of a composite. Before I do that, however, I want to explore the extent to which modal interpretations accomplish their purported goal to restore the empirical adequacy of the quantum theory. An argument by David Albert and Barry Loewer denies that modal interpretations accomplish this goal at all.

3.2.2 Albert and Loewer's Criticism

Albert and Loewer have presented their argument against the Kochen-Healey-Dieks (KHD) modal interpretation in several forms[21], but they all seem to fall under the following general schema:

[21] See Albert [1], and Albert and Loewer [2], [3] and [4].

(1) Ideal interactions have "measure-zero" in the set of all conceivable quantum interactions.

(2) Measure-zero implies physical impossibility.

(3) Thus, physically possible measurements are not ideal.

(4) But, the modal interpretation requires measurement interactions to be ideal.

(5) Hence, the modal interpretation cannot account for physically possible measurements.

The conclusion of the argument is that, under the KHD modal interpretation, quantum theory is empirically inadequate. Hence the goal that we set out to achieve, namely to restore the empirical adequacy of the theory by relinquishing the *e/e link*, is not after all accomplished. However, I claim that Albert and Loewer's argument, as it stands, cannot refute the modal interpretation. The problem concerns premises (1) and (2). These premises are used to derive statement (3), which is then employed as ammunition against the modal interpretation. What would make this part of the argument sound? First, the conclusion (3) must follow from the premises (1 and 2) and second, the premises must be true –of the same states of affairs in the world! More poignantly: "measure-zero" must mean the same in (1) and (2). But not all measures will indicate physical possibility. For instance, it can be shown that the set of states of the composite system that result from a non-ideal interaction are nondense in the norm topology of the tensor-product Hilbert space. However there are many well-known counterexamples to the claim that nondensity is a good criterion for physical impossibility. If "measure-zero" means "nondense in some norm topology" then premise (1) is true, but premise (2) is false in general, and hence the overall argument would seem to be unsound.

Albert himself has been trying to put this kind of reasoning to work in classical thermodynamics. The issue there concerns trajectories of thermodynamic states in phase space. Some of these trajectories represent systems

with increasing entropy, some with decreasing entropy. In the neighbourhood of every point in an entropy decreasing trajectory there is an arbitrary large number of points that belong to trajectories that represent entropy-increasing systems. But as Albert has pointed out[22], one cannot immediately draw inferences from this feature of phase space to the likelihood of those trajectories in the real world; instead one needs to investigate the thermodynamic systems, "*case by case*". Similarly in the quantum domain, Albert and Loewer need to produce specific examples of non-ideal interactions to support their conclusion. A measure-zero argument, on its own, will not do the job.

3.2.3 Non-ideal Measurements

The modal interpretation is committed to ideal interactions because it employs the biorthonormal decomposition rule to 'pick out' the observables that take values. At the end of the interaction, different values of the pointer position observable will be perfectly correlated with values of a well-defined observable over the object system; in other words there are no cross terms in the biorthonormal decomposition:

$$|\Psi\rangle = \sum_i d_i |\phi_i\rangle \otimes |\xi_i\rangle \tag{3.2.4}$$

But notice that this is a small subset of the set of all possible final states:

$$|\Psi\rangle = \sum_{ij} d_{ij} |\phi_i\rangle \otimes |\xi_j\rangle \tag{3.2.5}$$

For most states in this larger set, perfect correlations will no longer occur between values of the pointer position observable and the observable over the object system. One can put this in terms of the Hamiltonians that govern the dynamics of the composite. Ideal interaction Hamiltonians will yield final states that are already in the correct biorthonormal form (3.2.4). For the larger class of Hamiltonians that govern non-ideal interactions, the form of the final state of the combined system will be (3.2.5).

[22]In his lecture at the 1995 IUHPS meeting in Florence.

Notice that so far the modal interpretation is not really under threat, as its advocates could insist that it is possible in principle to give a treatment in terms of ideal interactions to all interesting cases of real physical interactions. However, Albert and Loewer continue:

> ...real measurements are almost never perfect in this sense. In a real measurement there is always some probability of the measuring device making an error. (Albert and Loewer, [3, p. 95]).

> If the world is anything less than entirely perfect (and of course it invariably is less than that), then the KHD interpretations don't end up doing their job right. And that's that. (Albert, [1, page 196]).

Albert and Loewer's criticism is that ideal measurement of the kind required by the modal interpretation are *physically impossible.* To support this claim they have invoked an argument from measure-zero:

> In the neighborhood of every Hamiltonian that characterizes an ideal measurement, there are Hamiltonians that characterize evolutions like [3.2.5]. In fact, on natural measures the measure of the set of Hamiltonians which correspond to ideal measurements is 0. (Albert and Loewer [4, p. 301]).

However, measure-zero rarely indicates physical impossibility. The classic counterexample is pointing a highly idealised one-dimensional pointer onto the real line. A natural measure can be laid out over the real line, on the basis of indifference, in such a way that every individual point will have measure-zero, while the whole line will have measure-one. But when the pointer points it will do so on one of those measure-zero points! One can construct highly counterintuitive examples. For instance, the classical mechanical model of the solar system yields a number of possible orbits, given some initial conditions. A measure *could be* laid out over the possible orbits of the planets (in the model), according to which the set of the nine

actual orbits will have measure-zero in the set of all possible orbits. And yet nobody will expect to be told that those orbits are impossible, or that they are '*almost never*' realised.

What sort of measure will do the job? And what will the measure be over? There are three possibilities: a measure over Hamiltonians, a measure over composite states, or a measure over the states of the subsystems.

Consider first a measure over the Hamiltonians that govern the evolution of the composite state. It is not clear that the set of Hamiltonians that induce ideal interactions must have measure-zero in the set of all Hamiltonians. Albert and Loewer certainly prove nothing of the sort; and even if the set of ideal Hamiltonians was to be given measure-zero, under some measure, the point made previously still holds: physical impossibility does not necessarily follow.

Secondly, consider a measure over the composite state. Recall that at the conclusion of a non-ideal measurement interaction the state of the composite is: $|\Psi\rangle = \sum_{ij} d_{ij}|\phi_i\rangle \otimes |\xi_j\rangle$, while if the measurement is ideal, we have: $|\Psi'\rangle = \sum_i d_i|\phi_i\rangle\otimes|\xi_i\rangle$. Every vector in the tensor product Hilbert space will represent a possible final state of the composite system. The advantage of focusing upon the final state of the composite, rather than the Hamiltonian, is that we can now make use of the metrical properties of the tensor-product Hilbert space to define the topological notion of density. Relative to some metric function d, a set S is dense if for any point $x \in S$ there is at least one point $y \in S$ such that $d(x, y) < \epsilon$, for arbitrarily small ϵ. Let us refer to those vectors in the tensor-product Hilbert space that result from ideal interactions as 'ideal state points', and those that result from non-ideal interactions as 'non-ideal state points'. The standard metric on the Hilbert space is defined by the norm operation: $d(x, y) = |(x - y)|$. It can then be shown that the set of ideal state points is non-dense in the topology induced by the norm: given a sufficiently small ϵ, the neighbourhood of any ideal state point will contain non-ideal state points only. The claim would then be that ideal state points are *unstable*, in the following sense: any arbitrarily small disturbance will take an ideal state point into a non-ideal state point in

its neighbourhood. Thus any small perturbation in the initial Hamiltonian will result in a non-ideal state point at the conclusion of the interaction; and, of course, some perturbation can never be totally excluded.

There are two obvious problems with this argument. First, it is far from clear exactly how 'disturbances' of the final states of the composite correspond to physical perturbations of the Hamiltonian. Why should tiny physical influences invariably get represented by small 'disturbances' of the final states? Second, there are many other metrics one could impose on the Hilbert space, in place of the norm-induced one. Under some of these metrics, the set of ideal state points will be dense. Is the norm-induced metric more justified than any other?[23]. The argument from instability would have more bite if the norm-induced metric of Hilbert space represented actual distances between objects in physical space. We could then intuitively see how a small force impressed upon the pointer of the apparatus may be represented by a small 'disturbance' of its state. But in fact, in the discrete model for measurements treated here, the Hilbert space norm is not intended to represent distances in space or spacetime. Instead, on the standard understanding of the theory, the norm has been associated with *probabilities* for outcomes of particular measurements in particular states[24].

A third option for Albert and Loewer is to somehow try to connect their measure to the quantum probabilities themselves. One way would

[23]In a manuscript recently brought to my attention, Lawrence Sklar raises a similar worry for the usual metric of distance in classical statistical phase space. It is clear that Sklar thinks that this is a most pressing and unresolved matter. He writes:

> The topologies that are usually invoked are, of course, those that are defined by using the usual metric of distance between points in the phase space. There will be many other metrics that will give rise to the same topology, but others, weird and deviant ones, that do not. But is our choice of the standard distance measure between the phase space points itself any more forced upon us by the dynamical laws and the constitution of the system [...]? (Sklar, [122, page 19])

[24]As in Born's probability rule (see Appendix 1).

be to define joint probabilities over the value states of the subsystems: $P_{ij}(|\phi_i\rangle\&|\xi_j\rangle) = \sum_l d_{ijl}$, where l ranges over all possible final states of the composite. (This probability would of course have to be conveniently normalised.) However, this proposal would not serve Albert and Loewer's purposes, because this measure cannot tell us that ideal interactions have measure zero. The joint probability for $|\phi_i\rangle$ and $|\xi_i\rangle$ is certainly not zero: the state of the composite at the conclusion of a non-ideal interaction is $|\Psi\rangle = \sum_{ij} d_{ij}|\phi_i\rangle \otimes |\xi_j\rangle$, and the probability of $|\phi_i\rangle\&|\xi_i\rangle$ is $|d_{ii}|^2$ which will typically be non-zero, and in fact should be quite large (how large depending upon the efficiency of the coupling interaction). So, on the basis of this measure we could not establish that the set of ideal interactions has measure zero.

In addition, this third proposal would have a further unintended consequence. The joint measure cannot be defined over the reduced states W_1 and W_2, as the information about the correlations contained in the composite state required by our measure is just the kind of information that gets lost when the reduced states are derived. Ultimately we would have to define Albert and Loewer's measure over pairs of value states, one of each subsystem. If the value states are given by the sets $\{|\phi_i\rangle\}$ and $\{|\xi_j\rangle\}$, we would then ascribe probabilities to all the pairs $\{|\phi_i\rangle, |\xi_j\rangle\}$. But, what would these probabilities be *of*? The simplest answer is: they would be probabilities for the actual states of the two subsystems at the conclusion of an randomly chosen measurement interaction. And indeed, by calculating the marginals of these joint probabilities we can define probabilities for the set of states $|\phi_i\rangle$ in W_1, and the set of states $|\xi_i\rangle$ in W_2.

3.2.4 The Ignorance Interpretation of Reduced States

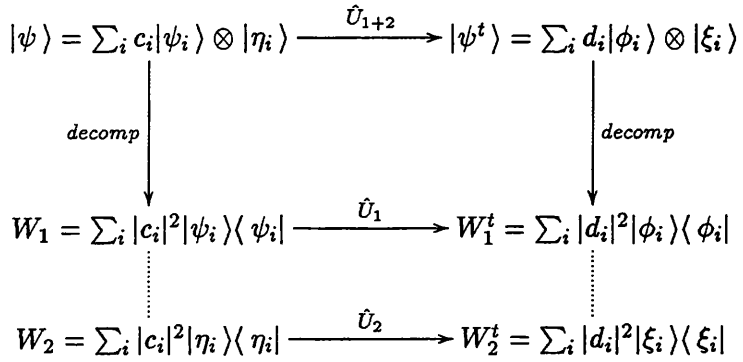
We must now finally address the following important question: Can we then apply the ignorance interpretation to the reduced states W_1 and W_2? Recall from section 3.1.2 that there are two arguments against the ignorance interpretation of mixtures, and one in its favour. First note that the main argument against clearly applies: W_1 and W_2 do not correspond to any

process of preparation, but are instead derived from the state of the composite by partial tracing; and the ignorance interpretation cannot generally be ascribed to improper mixtures.

I shall argue that, moreover, in this particular case the argument in favour of the ignorance interpretation of W_1 and W_2 is inapplicable. Recall that the argument in favour of the ignorance interpretation is its consistency with the time evolution of mixed states. But in this case the reduced states are not really *quantum mechanical* states: they do not evolve according to the Schrödinger equation. Here I shall give a shortened version of an argument to show that there is no consistent ascription of quantum mechanical states to the component subsystems, if those subsystems are really interacting[25]. Hence, I shall conclude, the reduced states W_1 and W_2 are not ignorance interpretable.

Suppose that Ψ is the state of the composite at the beginning of the interaction. Suppose further that W_1, W_2 are the reduced states of the subsystems at that same time. It may now seem natural to assume that the dynamics of the reduced states is given by the unitary operators U_1, U_2 acting on H_1, H_2; while the dynamics of the composite state is given by U_{1+2} acting on $H_1 \otimes H_2$. That is, we may try to complete our original diagram as follows:

[25]This is *not* to say that the *value states* have no Schrödinger evolution. It is as matter of course that value states can have no Schrödinger evolution; if they did evolve unitarily, the modal interpretation would lack the resources to do justice to the probabilistic character of quantum theory. The precise dynamics of value states is currently an area of some intense research (see for instance Bacciagaluppi and Dickson [6]).



However, such completion turns out to be impossible. Let me begin with stating an assumption. If the physics of the particular case at hand indicates that two systems evolve as free particles and have never interacted, nothing is gained in describing the dynamics of the composite. Conversely if the composite is ascribed an entangled state then we must take seriously the fact that the evolution of this composite state will exhibit the results of some physical interaction between the subsystems. One can then write the following criterion:

Criterion 3.2.1 *If two currently interacting systems are represented as subsystems in a composite system then the dynamics of the composite state will be given by a unitary operator $\hat{U}_{1+2}$ that is not factorizable in terms of independent operators $\hat{U}_1, \hat{U}_2$ that govern the evolution of the states of the component subsystems $\hat{U}_{1+2}|\Psi\rangle \neq \hat{U}_1 \otimes \hat{U}_2|\Psi\rangle$.*

Criterion 3.2.1 does not apply to systems that have interacted in the past but are currently not interacting -those systems might well show entanglement, while their current Hamiltonian does factorize. The logical form of the criterion is a conditional where the antecedent requires that the systems are *currently* interacting. The unitary operator that governs the evolution of this interacting dynamical state will not be factorizable in terms of operators acting independently on the states of the component subsystems. Otherwise we have not taken seriously the physics of the situation.

Now, the following lemma can be proved[26]:

Lemma 1 *If* $W_1^t = \hat{U}_1 W_1 \hat{U}_1^{-1}$ *and* $W_2^t = \hat{U}_2 W_2 \hat{U}_2^{-1}$, *where* $\hat{U}_1$, $\hat{U}_2$ *are unitary operators, then* $\hat{U}_{1+2}|\Psi\rangle = \hat{U}_1 \otimes \hat{U}_2|\Psi\rangle$

In other words it can be shown that if the dynamics of the reduced states is unitary, the unitary operator that governs the evolution of the dynamical state that represents the composite is factorizable[27]. But according to criterion (3.2.1) the state of the composite does **not** factorize in genuine interactions. It follows that in general the reduced states have no unitary evolution. Hence no consistent ascription of quantum states can be given to the individual subsystems in interaction. One might want to try to ascribe classical states instead, but classical states have a dynamics of their own, one that will generally be inconsistent with the the Schrödinger evolution of the composite state.

This result prevents us from applying in this case the one argument in favour of the ignorance interpretation of mixtures in general, namely its consistency with the *time-evolution* of mixed states. For in this particular case there is no time-evolution to show consistency with respect to.

3.2.5 Conclusions

I hope to have achieved two things. First I hope to have preempted the only argument that we may have initially entertained for actually giving the ignorance interpretation to the reduced states W_1, W_2, namely consistency

[26]The full result is in [106], but see Appendix 4 for a schematic proof. Brown's principle of Real Unitary Evolution (RUE) is crucially assumed in this proof. This is necessary if we don't want to preempt from the start the possibility that W_1 and W_2 be ignorance interpretable. For the ignorance interpretation entails (RUE).

[27]The lemma is true up to phase factors of the composite state. Under no circumstances can these phase factors alter the correlations between the subsystems. Of course an interaction with a further measuring device could be set up to reveal these factors, but the lemma would then apply to the subsystem involved in this further interaction.

with the time-evolution. This result vindicates Van Fraassen's cautionary remark: it is indeed only *as if* the ignorance interpretation of mixtures was correct; for in fact, it is not correct at all. The modal interpretation makes it possible to ascribe values to observables even when the subsystems are not in eigenstates: that is not state-ascription, it's value-ascription.

Secondly, I hope to have shown that Albert and Loewer's argument against KHD modal interpretations needs further empirical justification. A measure theoretic argument will not do, at least not without some further physical justification for the measure in question. Albert has suggested[28] that what underlies the intuition that real measurements are never ideal is 'common sense'. That seems right, provided that we understand such 'common sense intuition' as the result of accumulated knowledge of instances of real laboratory interactions, and abandon the attempt to ground the intuition directly upon the abstract structure of Hilbert space. Albert now agrees. The claim that ideal measurements are physically impossible, or very unlikely, cannot be justified by appeal to a dubious inference from a measure-zero set of states in Hilbert space. Such common sense intuition must be, if at all, supported by displaying actual instances of non-ideal interactions. In this context, that is work that still has to be done.

I am nevertheless inclined to go along with Albert and Loewer in the end, and agree with their claim that there exist irreducibly non-ideal real laboratory interactions[29]. But any argument for such a conclusion necessarily has to give a description of at least one such real laboratory interaction, and a plausible analysis of why the specific interaction must be modelled by means of a non-ideal interaction, and not, for instance, by means of an ideal measurement of an approximate observable.

[28] In conversation, Florence 1995.

[29] I argue that destructive interactions are conceivably of just this kind, in a joint paper with Marco DelSeta [43].

3.3 Measurement and Application

I have so far considered the problem of measurement as a threat against the empirical adequacy of the quantum theory. I call this the first measurement problem: is quantum theory *empirically adequate* of measurement interaction phenomena? The insolubility proof indicates that it is not, at least according to the standard interpretation of the theory. In order to solve the first problem of measurement we are requested to abandon the standard interpretational rule, the *generalised e/e link*, and to search for a new interpretation of quantum theory. The first problem of measurement is a problem in *interpretation*: what could it mean to say that a quantum system has, or possesses, a particular property, especially in cases when the system is in a superposition of eigenstates of the specific observable? The modal interpretation is an attempt to answer this question in full generality and without falling prey to the insolubility proof, and the Kochen-Specker proofs. In the previous section, I explored some of the main difficulties faced by the modal interpretation. Is the modal interpretation capable of fully rescuing the empirical adequacy of the quantum theory? The jury is, arguably, still out. It is still an open question whether the first problem of measurement can be solved or not.

Nevertheless, it is also possible to conceive the measurement problem in a different light, as a problem in the *application* of quantum theory. I call this the second measurement problem: can quantum theory be *applied to* measurement phenomena? When conceived in this manner the problem of measurement takes a rather different form; as a matter of fact the second problem of measurement *has* a solution and, I shall argue, the correctness of this solution does not in any way depend upon there being a solution to the first measurement problem.

3.3.1 Selective Interactions

In a series of recent papers [64], [66] [67], Arthur Fine has proposed a solution to the second measurement problem in terms of *selective interactions*.

Fine's approach makes use of an essential feature of the generalised formalism of quantum mechanics in terms of statistical operators, namely that it treats mixed and pure states on a par. Pure states are simply those states represented by idempotent statistical operators of trace one[30], i.e. they are just a subset among all possible states. The trace operation defines, for each state, pure or mixed, a statistical distribution over the eigenvalues of a particular observable. It becomes possible to group states, pure or mixed, into equivalence classes associated with every observable, depending on whether the states yield the same statistical distribution over the observable's eigenvalues.

Consider the definition of Q-equivalence (definition (4), in section 3.1.3): two states W and W' are Q-equivalent if and only if

$$prob(W, \hat{Q}) = prob(W', \hat{Q}),$$

i.e. if and only if both states lay out identical probability distributions over the eigenvalues of Q. If two states are Q-equivalent, their probability distributions are identical, and we call such states Q-indistinguishable. Thus, every observable Q effectively determines, for each state W, an equivalence class formed by all of W's Q-indistinguishable states. We can define the equivalence class $[W]_Q$ as follows:

Definition 6 (Q-Equivalence Class) $W \in [W]_Q$ *if and only if* $\forall W' \in [W]_Q : W \equiv_Q W'$

With this definition in mind, let us now return to the discussion of measurement interactions in section 3.1. A quantum object in state W_o interacts with an apparatus initially in state W_a. We are interested in the property O of the object, represented by the hermitian operator $\hat{O}$, with eigenvalues λ_i and eigenvectors ϕ_i. The pointer position observable on the apparatus is represented by the hermitian operator $\hat{I} \otimes \hat{A}$, with eigenvalues μ_{ni} and eigenvectors β_{ni}. The insolubility proof of the measurement problem shows that

[30] See Appendix 2, section 3.6.

no interaction can be set up that transfers the probability distribution laid out by W_o over the λ_i eigenstates of $\hat{O}$ to the μ_{ni} eigenstates of the pointer position observable, *if we allow that the initial state of the object may be any arbitrary state* including, crucially, superpositions of $\hat{O}$'s eigenstates.

Fine's proposal is to restrict the class of initial states of the object, according to the kind of observation that we are interested in. If the initial state of the object is W_o, and if we are interested in the observable O, Fine suggests that we focus only on the O-equivalence class of W_o, namely $[W_o]_O$; he advises, for the purposes of this particular interaction, (1) to ignore all other O-equivalence classes, formed by attending to states other than W_o, and (2) to disregard all X-equivalence classes of W_o for observables X of the object system other than O (i.e. those classes formed by considering the probability distribution that W_o would lay out over eigenstates of different observables). Suppose that O is a (discrete and not maximally degenerate) observable with spectral decomposition given by $\sum_n \lambda_n \hat{P}_n$, where $\hat{P}_n = \hat{P}_{[\phi_n]} = |\,\phi_n\rangle\langle\,\phi_n\,|$. We can construct the *standard representative* of $[W_o]_O$ as follows:

Definition 7 (Standard Representative) *$W_o(O)$ is the standard representative of the equivalence class $[W_o]_O$ if and only if:*

$$W_o(O) = \sum_n Tr(W_o \hat{P}_n) W_n = \sum_n prob_{W_o}(O = \lambda_n) W_n,$$

where $W_n = \left(1/Tr(\hat{P}_n)\right) \hat{P}_n$.

To construct the standard representative of an equivalence class, knowledge is required of both the initial state of a system, and of the spectral decomposition of a particular observable. Given such knowledge, the definition of the standard representative (definition 7) uniquely picks out a special state, typically a mixture, corresponding to each equivalence class. Moreover, this is a very special state indeed: it is a mixture over pure states for which the observable O takes some value with certainty. W_n is a normalised projector in the spectral decomposition of $\hat{O}$. Hence in each W_n the observable O will

take some value with probability one:

$$\forall W_n : prob_{W_n}(O = \lambda_n) = Tr(\hat{P}_n W_n) = 1.$$

This is, of course, easier to appreciate in the case of a maximal observable. For suppose that $\hat{O}$ has no degeneracies; then $W_n = \hat{P}_n$, and the probability for O to have a value λ_n in W_n is, simply: $prob_{W_n}(O = \lambda_n) = Tr(\hat{P}_n W_n) = Tr(\hat{P}_n \hat{P}_n) = Tr(\hat{P}_n) = 1$.

The standard representative $W_o(O)$ contains all the information relevant to measurements of the observable O in an object system initially in state W_o. Hence if an interaction is set up between the object system and the measurement apparatus to reveal the value of the property O in the object system, we need only be concerned, in modelling the interaction with the measuring apparatus, with the standard representative $W_o(O)$, not with the full state W_o. The interaction is 'selective', in the sense that it pays attention only to those features of the object system that are significant from the point of view of the relevant observable, while ignoring the remaining features concerning other observables.

Consider the evolution of the initial state of the composite. It is given by a unitary operator U_t:

$$U_t : W_o \otimes W_a \rightarrow U_t(W_o \otimes W_a)U_t^{-1}$$

The insolubility proof shows that, for an arbitrary state W_o, no such unitary operator exists that conforms to both the Transfer of Probability Condition (TPC) and the Occurrence of Outcomes Condition (OOC). The standard representative $W_o(O)$ is, however, not *any* arbitrary state. Fine's proposal is to replace the original W_o by the standard representative $W_o(O)$ of its equivalence class with respect to the relevant observable O. We then obtain[31]:

$$W_o(O) \otimes W_a \longrightarrow \qquad U_t(W_o(O) \otimes W_a)U_t^{-1} =$$

[31]The derivation below is for the simplest case of a maximal observable –the general case involves a further normalisation constant.

$$U_t(\sum_n Tr(W_o\hat{P}_{[\phi_n]})W_n \otimes \sum_n w_m\hat{P}_{[\gamma_m]})U_t^{-1} =$$
$$\sum_{nm} Tr(W_o\hat{P}_{[\phi_n]})w_m U_t(W_n \otimes \hat{P}_{[\gamma_m]})U_t^{-1} =$$
$$\sum_{nm} Tr(W_o\hat{P}_{[\phi_n]})w_m U_t(\hat{P}_{[\phi_n]} \otimes \hat{P}_{[\gamma_m]})U_t^{-1}$$

An interaction can easily be set up that effects the transfer of the entire probability distribution over O's eigenstates into a probability distribution over the eigenstates μ_{nm} of $I \otimes A$. Consider, for simplicity, an ideal non-disturbing interaction $U_t : U_t(\phi_n \otimes \gamma_m) = \phi_n \otimes \gamma_n$. Now, assuming (RUE), this interaction has the following effect:

$$\hat{P}_{[\phi_n]} \otimes \hat{P}_{[\gamma_m]} \longrightarrow U_t(\hat{P}_{[\phi_n]} \otimes \hat{P}_{[\gamma_m]})U_t^{-1} =$$
$$U_t(\hat{P}_{[\phi_n \otimes \gamma_m]})U_t^{-1} =$$
$$\hat{P}_{[\phi_n \otimes \gamma_n]} = \hat{P}_{[\beta_{nn}]},$$

where β_{nn} is an eigenvector of $(\hat{I} \otimes \hat{A})$ with eigenvalue μ_{nn}. The final state of the composite is then:

$$W^f_{o+a} = \sum_{nm} Tr(W_o\hat{P}_n)w_m\hat{P}_{[\beta_{nn}]} = \sum_{nm} \eta_{nm}\hat{P}_{[\beta_{nn}]}. \qquad (3.3.6)$$

This is a mixture over pure states, namely projectors associated with the eigenvectors of $I \otimes A$. According to part (b) of the definition of the *generalised e/e link* (definition 3), we are entitled to say that in this final state of the composite $I \otimes A$ has a value, in accordance with (OOC).

In other words, by restricting the initial state to its standard representative it becomes possible to escape the insolubility proof of the quantum measurement problem. It now becomes possible to set up a Schrödinger interaction that obeys *both* the Transfer of Probability Condition (TPC) and the Occurrence of Outcomes Condition (OOC) –without ever relinquishing any standard interpretational rule, such as the *e/e link*.

3.3.2 Ignorance, and State-descriptions

I claim that Fine's proposed way around the insolubility proof constitutes a solution to the second measurement problem, concerned with the *application*, rather than empirical adequacy, of the quantum theory. In any case

where the application of a theory is not guided by a concern to test the theory, we expect an independently standing mediating model to be doing much of the work, otherwise normally done by the theory itself, in fixing the criteria for the introduction of corrections and amendments into the theoretical descriptions. I shall argue that the mediating model employed in Fine's proposal is a conception of a probability distribution over the eigenvalues of an observable as a *physical aspect* of an object. An appreciation of this conception can be gained by considering the distinction between state-ascription and state-description, and the important function that the latter plays in Fine's proposal.

Let me then explore a bit further the sense in which 'selective' interactions constitute a solution to the measurement problem. The result of a selective interaction is a weighted sum over states in each of which the pointer position observable has values:

$$W^f_{o+a} = \sum_{nm} \eta_{nm} \hat{P}_{[\beta_{nn}]}.$$

The final state is mixed, and we are advised by the conjunction of (OOC) and the *generalised e/e link*, to give this mixture the ignorance interpretation. That is, we are asked to infer that the state of the composite *really* is one of the states $\hat{P}_{[\beta_{nn}]}$, although we ignore precisely which one it is; one of these states *really* obtains at the end of the interaction with the prescribed probability η_{nm}. This ignorance interpretation of W^f_{o+a} is indeed possible: the mixture in question is not an improper one, as it has not been derived from the state of a larger composite system by employing the standard identifications (3.7.11). As a matter of fact the (*object + apparatus*) system itself need not be conceived, for our purposes, as a subsystem of any larger physical system. On the contrary, it is natural, on this picture, to think of the (*object + apparatus*) system as a closed system. The final state of the composite, W^f_{o+a} is a proper mixture, for it is the result of a preparation procedure –a complex and highly convoluted procedure perhaps, but a preparation procedure nevertheless. It should then be possible to give W^f_{o+a} the ignorance interpretation, as required by (OOC).

It is now worth looking into the origin of the probabilities η_{nm} that appear in W^f_{1+2}. For instance η_{11} (the probability for the β_{11} state in this mixed state, which we are supposed to interpret as a measure of subjective ignorance) is given by the product $Tr(W_o\hat{P}_1)\times w_1$. Now, w_1 is the probability of the pure state $\hat{P}_{[\gamma_1]}$ in the initial apparatus mixture[32], while $Tr(W_o\hat{P}_1)$ is the probability ascribed to the pure state W_1 in the standard representative of the object system. We are asked, in Fine's proposal, to give the final state of the composite an ignorance interpretation, and to understand the probabilities involved in the state W^f_{o+a} as subjective probabilities describing our incomplete knowledge of the 'true' state. But it is now clear that in order to do that, we have to equally give the mixtures $W_o(O)$ and W_a an ignorance interpretation. For the probability η_{11} ascribed in W^f_{o+a} to the state β_{11} is a product of two probabilities, w_1 and $Tr(W_o\hat{P}_1)$. The subjective interpretation of η_{11} requires the subjective interpretation of both w_1 and $Tr(W_o\hat{P}_1)$. Hence, in order to give the ignorance interpretation to the final state of the composite we have to give the ignorance interpretation to both the initial state of the apparatus W_a, and the standard representative of the system, $W_o(O)$.

This is again not incompatible with the arguments against the ignorance interpretation of improper mixtures that I reviewed in section 3.1.2. For neither W_a nor $W_o(O)$ is an improper mixture, as neither is found by 'derivation' from states of larger systems that the composite (*object + apparatus*) may be a physical part of.

But to give an ignorance interpretation to $W_o(O)$ raises a puzzle. $W_o(O)$ is a mixture over W_n states:

$$W_o(O) = \sum_n Tr(W_o\hat{P}_n)W_n = \sum_n prob_{W_o}(O = \lambda_n)W_n.$$

[32]Note, *en passant*, that the actual composition of W_a has no effect whatever on the final state of the composite. This is as it should be for an ideal interaction, which is the only case I treat here. In general, however, the initial state of the apparatus *will matter*. For that reason only I have kept throughout these formulae reference to the index m which, in the special case of an ideal interaction, becomes just a dummy index.

In giving the ignorance interpretation to $W_o(O)$ we will be claiming that the *true* state of the object system, at the beginning of the interaction, is *really* one of the states W_n, with the prescribed probabilities. But, the *true* initial state of the system is W_o! This may not even be a mixed state, and it will generally be a very different state to any of W_n. Moreover, although by construction the mixture $W_o(O)$ is in W_o's O-equivalence class, neither of the pure states W_n that appear in the decomposition of $W_o(O)$ is.

The point can be made more graphically by considering the very case that gave rise to the insolubility proof. There, we were invited to consider a two-dimensional observable O with eigenstates ϕ_1 and ϕ_2 and corresponding eigenvalues λ_1 and λ_2. We were then asked to consider three O-distinguishable states, ϕ_1, ϕ_2 and ϕ_3, where ϕ_3 was the linear combination: $\phi_3 = a_1\phi_1 + a_2\phi_2$. Given ϕ_3 and the spectral decomposition of $\hat{O} = \lambda_1\hat{P}_{\phi_1} + \lambda_2\hat{P}_{\phi_2}$ we can construct the standard representative in ϕ_3's O-equivalent class, namely the mixed state: $W_o(O) = |a_1|^2|\,\phi_1\,\rangle\langle\,\phi_1\,| + |a_2|^2|\,\phi_2\,\rangle\langle\,\phi_2\,|$.

Fine's proposal demands that we give the ignorance interpretation to $W_o(O)$. Giving the ignorance interpretation to $W_o(O)$ amounts to claiming that the system is *really* in state ϕ_1 with probability $|a_1|^2$, or in state ϕ_2 with probability $|a_2|^2$. But we know that the state of the system is really neither ϕ_1, nor ϕ_2, but ϕ_3, which is a superposition of both. Surely we are not here being asked to entertain the long-discredited ignorance interpretation of superpositions[33]. What is going on?

What is going on, I think, is that we are implicitly abandoning the requirement that systems have, or possess, states *really*. Instead, systems have state-descriptions. Descriptions are useful as long as they enable us to solve a particular problem. State-descriptions that are useful for some purposes may not be useful for other purposes. The description of the initial object system as being in the superposition helps us to understand typical quantum interference phenomena that the system may generate. But, as

[33]The irony won't be lost to the reader that this is precisely the case of Schrödinger's cat (one only has to think of ϕ_1 and ϕ_2 as the dead/alive eigenstates).

the insolubility proof makes manifest, it does not help at all to understand why measurements of properties of this system have outcomes. For that we need to change the description, and adopt the corresponding mixture, appropriately ignorance-interpreted.

3.3.3 Equivalence Classes as Physical Aspects

Although I believe that Fine's proposal presupposes *some* form of instrumentalism towards physical theory, I do not think that it is of the same kind that sometimes gets associated with the logical positivist tradition. In the early works of Nagel [107, pages 129-140] and Hempel [83, chapter 8], instrumentalism was understood as a thesis concerning the meaning of theoretical terms; it was an interpretation of theory. I want to use the term in a much broader sense, as an *attitude* towards science, and as an allied methodology for scientific knowledge. Neither do I see instrumentalism as the feeble and weak position that it is sometimes taken to be in contemporary discussions –a sort of 'anything goes' attitude that aims to 'debunk' the authority of science by revealing that its claims about the world, when taken literally, lack any true import. Instead I see instrumentalism as a strong, robust and positive approach to scientific knowledge, a *pro-science* attitude.

Here I can only give a preliminary sketch. This kind of instrumentalism is essentially a two-vector notion. One points towards empiricism, and against realism, and asks us to refrain from *projecting out* our representations onto the world. As Fine aptly puts it, the instrumentalist claims that *our descriptions of physical systems are 'in the head', and not out there in the world.* The second component, though, points away from empiricism, and towards pragmatism. The justification for the use of a particular representation, or a description, of a concrete system is to be found in this description's pragmatic virtues (convenience, tractability, simplicity, consistency with previous representations, instrumental reliability, and so on). Pragmatic justification is not always epistemic justification. There is no reason to think that pragmatic justification will always track an empiricist-oriented sense of epistemic justification, –one, for instance, of the kind that

Van Fraassen's constructive empiricist would wish to defend.

From this instrumentalist point of view, physical systems do not *possess* states, rather they have *state-descriptions.* The use of one particular description rather than another is justified by the success of the description in solving a particular problem at hand. This pattern of justification provides an effective procedure for application. The procedure for applying quantum theory, for instance, is as follows. If we are concerned with an experiment designed to exhibit the interference phenomena characteristic of a quantum superposition, we better use the full initial state of the object system W_o in modelling the phenomenon, for only that state can give us the correct interference results. If, on the other hand, the experiment is intended to measure a certain quantity O on a quantum system, by means of an interaction with a measuring device, we should make use of the standard representative $W_o(O)$; only by using this state will we succeed in modelling the measurement process fully within quantum mechanics.

For any interactive quantum system, the theory seems to yield two inconsistent future predictions, corresponding either to a measurement or an interference interaction. We find out which one applies by inspecting the experimental context, and we choose the description of the initial state of the system that is more appropriate. As a procedure to *test* the theory, this is clearly unsatisfactory, for it seems *ad hoc.* But this is precisely the point, *for the theory is not being put to the test.* In modelling measurements as selective interactions we are not concerned with finding out if the theory is true, or empirically adequate; we are only concerned with applying the theory to a particular phenomenon, namely the phenomenon of the occurrence of outcomes in quantum measurement interactions. And as Duhem stressed[34], "*there is nothing to shock logic in this procedure*".

Although the decision as to whether to use W_o, $W_o(O)$, or the standard representative of any other among W_o's X-equivalence classes is not dictated by the quantum theory (and hence seems *ad hoc* from the point of view of

[34]See the quotation in section 1 of chapter 2.

the theory), it is not an arbitrary decision. It is informed by a mediating model of the kind I described in chapter 1. Fine writes:

> The basic proposal, then, is to regard the measurement of an observable E on a system in state ψ as a measurement interaction that selects the aspect of the system corresponding to the probability distribution for E that is determined by state ψ. (Fine, [67, page 126]).

> The probability distributions pick out physical aspects that are sufficient to determine the patterns of outcomes of selective interactions. (Fine, [64, page 503]).

We are invited to conceive an *equivalence class* for some observable O and some state W_o of an object, as a *physical aspect* of the object. This conception is not forced upon us by the phenomena, nor is it part of any established theory. But it contains knowledge about the structure of quantum objects, which helps us to choose the appropriate description of the initial state of the system. Fine takes systems to relate to their aspects in the same way wholes relate to parts, and he uses the system-subsystem analogy:

> My exploration starts out from the idea that some interactions are selective. They do not actually involve the whole system, only some physical subsystem. Thus the interaction formalism ought not to be applied to the state of the whole system, only a representative of the subsystem engaged in the interaction. (Fine, [64, page 502]).

The conception of equivalence classes as physical aspects enables us to apply the quantum theory to measurement interactions. Recall that every specific probability distribution over the eigenvalues of some observable picks out an equivalence class of states; and for every equivalence class there corresponds a probability distribution for an observable. Every possible state of a system defines one such probability distribution, and a corresponding equivalence class, for <u>each</u> observable. (The state W_o defines a family of X-equivalence

classes, where X ranges over all the observables of the system.) It seems reasonable then, to demand that we treat an interaction of the system with a measuring apparatus, designed to reveal information about one particular observable of the system, as an interaction of the apparatus with *just that part* of the system.

The criteria for application of the quantum theory is in this case laid out by a mediating model, of just the sort described in chapter 1 of the Thesis. The presence of this model makes Fine's proposal a superior alternative, as an application of the quantum theory to measurement interactions, to some of its 'pragmatic' competitors. For instance, one could replace the *final* superposition with the correct mixture. Indeed this would seem no more *ad hoc*, from the point of view of quantum theory, than Fine's proposal to substitute the standard representative for the *initial* superposition. However, from the point of view of the mediating model, this would be a totally arbitrary move. It would be justified only as long as it can be shown to yield precisely the same mixture that would result from an interaction with the appropriate aspect of the system, –represented by the standard representative of the equivalence class defined by the initial object state and the relevant observable. In this case, the alternative proposal would be equivalent to Fine's for all practical purposes.

In the first part of the Thesis I argued that the structure of application and the structure of confirmation differ. Typically the domain of application of a theory is larger than its domain of empirical adequacy. Here, I think, we have an instance. The quantum theory *can be*, after all, applied to the phenomena of measurement interactions. We may well come to convince ourselves, *pace* the possible success of modal interpretations, that the quantum theory is empirically inadequate, precisely on the grounds that it cannot embed all measurement interaction phenomena. This may be so, but it constitutes no reason to think that a consistent *application* of quantum theory to measurement phenomena cannot be found.

3.4 Summary

In the first section, I introduced the problem of measurement as a threat to the empirical adequacy of the quantum theory. Some features of the interaction formalism of quantum mechanics seem to disagree as a matter of fact with some well confirmed facts about nature.

In the second section, I reviewed Van Fraassen's favourite proposal to dispel the threat, and to solve the measurement problem, by giving a particular interpretation to the quantum theory, the so-called *modal interpretation.* I focused on the version due to Kochen, Healey and Dieks, and I reviewed a recent critique of their interpretations, due to Albert and Loewer.

In the last section I discussed Fine's proposal to solve the measurement problem by finding alternative strategies to apply the quantum formalism. Fine construes the measurement problem as a problem in the application of quantum theory, rather than a problem for the theory's empirical adequacy. He does not attempt to provide an interpretation of the formalism, and focuses instead on deriving a useful model for quantum measurement interactions.

3.5 Appendix 1: Basic Principles of Quantum Theory

In this appendix I review the principles of quantum theory in a simple framework that assumes that the states of a system can be represented as vectors in Hilbert space. In this simplest form, quantum theory is roughly thought to consist in the following set of five principles[35]:

1. **States.** States of physical systems are represented by normalised vectors ($|\psi\rangle$ such that $|\langle\psi|\psi\rangle|^2 = 1$) in a Hilbert space H.

[35]These principles have to be altered to make room for *mixed states* (see Appendix 2, section 3.6).

2. **Observables.** The measurable quantities ('observables') of physical systems are represented by linear Hermitian operators ($\hat{A}$ defined as (1) linear: $\hat{A}(a(\vec{v}+\vec{w})) = a\hat{A}\vec{v} + a\hat{A}\vec{v}$, and (2) self-adjoint: $\langle \vec{v}|\hat{A}\vec{w}\rangle = \langle \hat{A}\vec{v}|\vec{w}\rangle$) that act on Hilbert space. Hermitian operators have real eigenvalues:

$$\langle \hat{A}\vec{v}|\vec{v}\rangle = a^*\langle v|v\rangle = \langle v|\hat{A}\vec{v}\rangle = \langle v|av\rangle = a\langle v|v\rangle \Longrightarrow a^* = a$$

Also, importantly, the eigenvectors of Hermitian operators are mutually orthogonal. If $\hat{A}|v\rangle = a_1|v\rangle$, and $\hat{A}\vec{w} = a_2|w\rangle$, then:

$$\begin{aligned}\langle \hat{A}\vec{v}|\vec{w}\rangle = \langle \vec{v}|\hat{A}\vec{w}\rangle = \langle a_1\vec{v}|\vec{w}\rangle = \langle \vec{v}|a_2\vec{w}\rangle = \\ = a_1\langle \vec{v}|\vec{w}\rangle = a_2\langle \vec{v}|\vec{w}\rangle \Longrightarrow \langle \vec{v}|\vec{w}\rangle = 0 \\ \Longrightarrow \vec{v} \perp \vec{w}\end{aligned}$$

The eigenstates of a maximal observable, represented by a non degenerate operator, form a complete basis that spans the whole space. Thus any vector state in the Hilbert space can be written as a linear combination of the eigenstates of any non degenerate operator.

3. **Dynamical Evolution.** The evolution of states is subject to the Schrödinger equation:

$$i\hbar\frac{d\psi}{dt} = H\psi \tag{3.5.7}$$

where H represents the Hamiltonian of the system.

There are several important aspects to the dynamics. First, the quantum mechanical evolution is continuous from one state ψ to the next ψ_t. Hence there must be some operator $\hat{U}_t$ acting on the space that takes from ψ to ψ_t. This is a complex function of the Hamiltonian:

$$\hat{U}_t = e^{-iHt} \tag{3.5.8}$$

$\hat{U}_t$ is a *unitary* operator. It is linear: $\hat{U}(a\vec{v}+\vec{w}) = a\hat{U}\vec{v}+a\hat{U}\vec{w}$; it has an inverse $\hat{U^{-1}}$ such that $\hat{U}\hat{U^{-1}} = \hat{U^{-1}}\hat{U} = \hat{I}$; and it is norm-preserving: $\forall\vec{v} : |\hat{U}\vec{v}|^2 = |\vec{v}|^2$.

Second, the linearity of the Schrödinger equation implies the principle of superposition for quantum states. If two vectors $\vec{v}$ and $\vec{w}$ in the Hilbert space represent possible states of a quantum system, then the superposition $\psi = \vec{v} + \vec{w}$ also represent a possible state of the system. Finally the unitaricity of the quantum mechanical evolution entails that no pure state ever evolves into a mixed state by Schrödinger evolution.

4. **Empirical Content.** The connection with observed outcomes in experiments is supposedly established by Born's probability rule. According to this rule, the probability that, on measuring observable A in a system in state ψ, the outcome a_i is found, is given by:

$$Prob_{\psi}(A = a_i) = \langle \psi | a_i \rangle \langle a_i | \psi \rangle = |\langle \psi | a_i \rangle|^2.$$

where $\{|a_i\rangle\}$ is the set of eigenvectors of $\hat{A}$ with eigenvalues $\{a_i\}$.

Similarly the *expectation value* of an observable A for a system in state ψ can be defined as follows: $Exp_{\psi}(A) = \langle \psi | \hat{A} | \psi \rangle$. This definition is consistent with Born's rule for the probability that a system takes a particular value of some observable. To see this, consider a system in state $\psi = \sum_j c_j |a_j\rangle$, where $\{|a_j\rangle\}$ is the set of eigenstates of the operator A with corresponding eigenvalues $\{a_j\}$. Then:

$$\langle \psi | \hat{A} | \psi \rangle = \sum_i a_i^* \langle a_i | \hat{A} \sum_j a_j |a_j\rangle = \sum_i a_i^* \langle a_i | \sum_j a_j \hat{A} |a_j\rangle =$$
$$\sum_i a_i^* \langle a_i | \sum_j a_j c_j |a_j\rangle = \sum_i \sum_j a_i^* a_j^* c_j \langle a_i | a_j \rangle =$$
$$\sum_i a_i^* a_i c_i = \sum_i c_i |a_i|^2 = \sum_i c_i Prob_{\psi}(A = a_i).$$

which is the expectation value, given Born's rule, as required. The rule $Exp_{\psi}(A) = \langle \psi | \hat{A} | \psi \rangle$ is also sometimes referred to as Born's probability rule. Born's probability rule can be expressed in a very convenient statistical algorithm in terms of projection operators which we now derive.

A projection operator is an idempotent ($\hat{P}^2 = \hat{P}$) Hermitian operator which maps every vector in the Hilbert space into its geometrical projection along a particular subspace in the space (a one-dimensional projector maps vectors into their geometrical projections along *rays* in Hilbert space). One can associate a projection operator with each eigenvalue of any particular observable. For example, suppose that $\hat{O}$ is non-degenerate and has $\{|\psi_i\rangle\}$ as its eigenstates, and $\{o_i\}$ as its eigenvalues. The family of projectors can be defined: $P_i = P_{[\psi_i]} = |\psi_i\rangle\langle\psi_i|$. Projector $P_1 = |\psi_1\rangle\langle\psi_1|$, for instance, takes any arbitrary vector $|\psi\rangle = \sum_i c_i|\psi_i\rangle$ in H into its one-dimensional subspace spanned by $|\psi_1\rangle$: $P_1|\psi\rangle = |\psi_1\rangle\langle\psi_1|\sum_i c_i|\psi_i\rangle = \sum_i c_i|\psi_1\rangle\langle\psi_1|\psi_i\rangle = c_1|\psi_1\rangle$.

We are now in a position to state an important theorem:

Theorem 3.5.1 (Spectral Decomposition Thm) *For every Hermitian operator $\hat{O}$ there is a set of scalars $\{o_1, ..., o_m\}$ and a set of projectors $\{\hat{P}_1, ..., \hat{P}_2\}$ projecting onto mutually orthogonal subspaces of H, such that:*

$$\hat{O} = \sum_i^m o_i \hat{P}_i = \sum_i^m o_i |\psi_i\rangle\langle\psi_i|.$$

There is always a spectral decomposition for any Hermitian operator. Moreover, if $\forall i \neq j : o_i \neq o_j$ then the spectral decomposition is unique (for the details see Redhead [115, page 13] or Hughes [89, page 50]). Our desired result, the Quantum Statistical Algorithm is the application of Born's rule to the spectral decomposition:

$$Prob_\psi(O = o_i) = \langle\psi|\hat{P}_i^O|\psi\rangle. \tag{3.5.9}$$

where $\hat{P}_i^O$ is in the spectral decomposition of $\hat{O}$ associated with eigenvalue o_i.

5. **Repeatability of Outcomes (or constancy of properties).** If observable A has been measured on a system in state ψ, and the outcome is a_i, then a subsequent measurement of this same observable

immediately after will yield value a_i again, with certainty ($Prob_\psi(A = a_i) = 1$). The obvious way to satisfy this requirement is by imposing the so-called *collapse postulate* and demanding that the system's state be transformed into the corresponding eigenstate $|a_i\rangle$. This 'instantaneous transition' $\psi \rightarrow \pm a_i$ cannot be consistent with Schrödinger evolution as it is a transition from a pure to a mixed state. Von Neumann [109, section V] introduced the concept of dual evolution precisely to account for this type of indeterministic transition.

3.6 Appendix 2: Mixed States and Statistical Operators

In this appendix I discuss a more general framework, in terms of statistical operators, that enables us to deal with systems in mixed states. Mixed states are represented by statistical operators: $W = \sum_i p_i \langle v_i | v_i \rangle$. Our aim is to derive an analogue of the quantum statistical algorithm for statistical operators. Suppose that the Hermitian operator $\hat{O}$ has spectral resolution $\sum_i \hat{P}^O_{o_i}$ where o_i ranges over the possible eigenvalues of $\hat{O}$. To each pure state $|v_i\rangle$ in W there corresponds an expectation value of O, namely $\langle v_i|\hat{O}|v_i\rangle = \sum_n o_n c_n^{(i)*} c_n^{(i)}$. The mean value of O in the mixture W is the sum of all these expectation values appropriately weighted by the corresponding $\{p_i\}$:

$$Exp_W(O) = \sum_i p_i \langle v_i|\hat{O}|v_i\rangle = \sum_n o_n \sum_i p_i c_n^{(i)*} c_n^{(i)}.$$

We define W's *density matrix* in the $\{|\psi_i\rangle\}$ basis as: $\rho_{nn} = \sum_i p_i c_n^{(i)*} c_n^{(i)}$. ($\rho_{nn}$ is the matrix representation of W in that basis). Hence:

$$Exp_W(O) = \sum_n o_n \rho_{nn} =$$

$$\sum_n O_{nn} \rho_{nn} = \sum_n (\hat{O}\hat{W})$$

where O_{nn} is the diagonal matrix of $\hat{O}$ in the basis $\{|\psi_i\rangle\}$. We define $Tr(\hat{O}\hat{W}) = \sum_n (\hat{O}\hat{W})$. The trace ($Tr$) operation on statistical operators satisfies, among other, the following conditions:

1. It is symmetric: $Tr(\hat{W}\hat{O}) = Tr(\hat{O}\hat{W})$.

2. Quantum states (mixed or pure) are represented by trace-one operators: $Tr(\hat{W}) = 1$. (This follows from the fact that in $\hat{W} = \sum_i p_i P_{[\Psi_i]}$, $\sum_i p_i = 1$).

3. If $\hat{W}$ represents a mixed state: $Tr(\hat{W}^2) < Tr(\hat{W})$. (As $0 \leq p_i \leq 1$, and hence $p_i^2 < p_i$ for all $i \neq 0, 1$).

4. If $\hat{W}$ represents a pure state: $Tr(\hat{W}^2) = Tr(\hat{W})$. (A pure state $|\Psi_n\rangle$ is represented by a projector $\hat{P}_{\Psi_n}$, an idempotent statistical operator. Hence for a pure state there is some n for which $p_i = 1$ if $i = n$, $= 0$ otherwise).

The Generalised Born Rule for statistical operators, which determines the expectation value of any observable O for a system in state W, takes the form:

$$Exp_W(O) = Tr(\hat{O}\hat{W})$$

Alternatively we define the probability for an observable O of a system in state $\hat{W}$ to take value o_i as

$$Prob_W(O = o_i) = Tr(\hat{W}\hat{P}_i^O). \tag{3.6.10}$$

where P_i^O is the projector associated with eigenvalue o_i in the spectral decomposition of O.

3.7 Appendix 3: The Interaction Formalism

In this appendix I describe the tensor-product space formalism that quantum theory provides to represent interactions between systems, which is essential in discussing the measurement problem. Given two Hilbert spaces H_1 and H_2 we can always form the *tensor-product* Hilbert space $H_{1+2} = H_1 \otimes H_2$. The dimension of the tensor-product space is the product of the dimensions of the individual spaces, $dim(H_{1+2}) = dim(H_1) \times dim(H_2)$. If $\{|v_i\rangle\}$ is

a basis for H_1 and $\{|w_j\rangle\}$ a basis for H_2 then $\{|v_i\rangle \otimes |w_j\rangle\}$ is a basis for H_{1+2}. Similarly if A is an observable defined on H_1 with eigenvectors $|v_i\rangle$ and eigenvalues a_i, and B an observable on H_2 with eigenvectors $|w_i\rangle$ and eigenvalues b_j, then $A \otimes B$ is an observable on H_{1+2} with eigenvectors $|v_i\rangle \otimes |w_j\rangle$ and corresponding eigenvalues $a_i b_j$.

Consider two systems S_1 and S_2. If S_1's state is W_1 on H_1, and S_2's state is W_2 on H_2, we can represent the state of the combined system S_{1+2} as the statistical operator $W_{1+2} = W_1 \otimes W_2$ acting on the tensor-product Hilbert space H_{1+2}. If either W_1, W_2 is a mixture, then W_{1+2} is also a mixture. If, on the other hand, *both* W_1, W_2 are pure states then W_{1+2} is pure. Suppose that $W_1 = P_{|\psi\rangle}$, and $W_2 = P_{|\phi\rangle}$, where $|\psi\rangle = \sum_i c_i|v_i\rangle$ and $|\phi\rangle = \sum_j d_j|w_j\rangle$, then $W_{1+2} = \sum_{i,j} c_i d_j |v_i\rangle \otimes |w_j\rangle$, which is a superposition of eigenstates of $A \otimes B$ in H_{1+2}. More specifically, if S_1, S_2 are in eigenstates of A, B, the combined system S_{1+2} is in an eigenstate of $A \otimes B$. If $W_1 = |v_i\rangle$ and $W_2 = |w_j\rangle$, then $W_{1+2} = |v_i\rangle \otimes |v_j\rangle$, a so-called *product state.*

For an arbitrary (pure or mixed) state W_{1+2} of the combined system, and arbitrary observable $A \otimes B$ the Generalised Born Rule applies. The probability that $A \otimes B$ takes a particular $a_i b_j$ value is given by:

$$prob_{W_{1+2}}(A \otimes B = a_i b_j) = Tr(\hat{W}_{1+2}\hat{P}_{ij}),$$

and the expectation value of the 'total' $A \otimes B$ observable in state W_{1+2} is:

$$Exp_{W_{1+2}}(A \otimes B) = Tr((\hat{A} \otimes \hat{B})\hat{W}_{1+2})$$

We will sometimes be given the state W_{1+2} of a composite system, and then asked to figure out what the *reduced* states W_1, W_2 of the separated subsystems must be. Given a couple of observables A and B on H_1, there are some relatively straightforward identifications that help to work out the reduced states, namely:

$$\begin{aligned} Tr((\hat{A} \otimes \hat{I})\hat{W}_{1+2}) &= Tr(\hat{A}\hat{W}_1) \\ Tr((\hat{I} \otimes \hat{B})\hat{W}_{1+2}) &= Tr(\hat{B}\hat{W}_2) \end{aligned} \qquad (3.7.11)$$

where $\hat{I}$ is the identity observable $(\forall|\psi\rangle : \hat{I}|\psi\rangle = |\psi\rangle)$. This amounts to the demand that the probability distribution set up by observable A (B) in

the reduced state W_1 (W_2) be the same as that laid out by $A \otimes I$ ($I \otimes B$) in the composite state W_{1+2}, –thus effectively ensuring that the choice of description (either in the larger or smaller Hilbert space) of a subsystem in a larger composite system, has no measurable consequences as regards the monadic properties of the individual subsystem.

3.8 Appendix 4: A Lemma for Reduced States

In this appendix I give a schematic proof of the lemma presented in section 3.2.4 (lemma 1).

Lemma 2 (1) *If* $W_1^t = \hat{U}_1 W_1 \hat{U}_1^{-1}$ *and* $W_2^t = \hat{U}_2 W_2 \hat{U}_2^{-1}$, *where* $\hat{U}_1$, $\hat{U}_2$ *are unitary operators, then* $\hat{U}_{1+2}|\Psi\rangle = \hat{U}_1 \otimes \hat{U}_2|\Psi\rangle$.

Consider the following diagram:

$$
\begin{array}{ccc}
|\psi\rangle = \sum_i c_i|\psi_i\rangle \otimes |\eta_i\rangle & \xrightarrow{\hat{U}_{1+2}?} & |\psi^t\rangle = \sum_i c_i|\phi_i\rangle \otimes |\xi_i\rangle \\
\Big\downarrow \textit{decomp} & & \Big\uparrow \textit{decomp} \\
W_1 = \sum_i |c_i|^2 |\psi_i\rangle\langle\psi_i| & \xrightarrow{\hat{U}_1} & W_1^t = \sum_i |c_i|^2 |\phi_i\rangle\langle\phi_i| \\
\vdots & & \vdots \\
W_2 = \sum_i |c_i|^2 |\eta_i\rangle\langle\eta_i| & \xrightarrow{\hat{U}_2} & W_2^t = \sum_i |c_i|^2 |\xi_i\rangle\langle\xi_i|
\end{array}
$$

The diagram assumes that the reduced states correspoding to the subsystems have their own independent unitary evolution and asks what are the constraints on the unitary evolution of the state of the composite. According to (RUE), the evolution of the subsystems leaves the coefficients $|c_i|^2$ invariant, and the evolved reduced states are:

$$
\begin{aligned}
W_1^t &= \sum_i |c_i|^2 \hat{U}_1 |\psi_i\rangle\langle\psi_i| \hat{U}_1^{-1} = \sum_i |c_i|^2 |\phi_i\rangle\langle\phi_i| \\
W_2^t &= \sum_i |c_i|^2 \hat{U}_2 |\eta_i\rangle\langle\eta_i| \hat{U}_2^{-1} = \sum_i |c_i|^2 |\xi_i\rangle\langle\xi_i|
\end{aligned}
$$

These must correspond to some state of the composite, from which they are derived by partial tracing. According to Appendix 3, this state is given as:

$$|\psi^t\rangle = \sum_i c_i |\phi_i\rangle \otimes |\xi_i\rangle = \sum_i c_i \hat{U}_1 |\psi_i\rangle \otimes \hat{U}_2 |\eta_i\rangle$$

which is already in a biorthonormal decomposition form. Note that the state of the composite is determined by the reduced states only up to a phase factor (and not necessarily a mere global phase factor, for c_i may take negative values in some but not all of the terms in the sum.) State $|\psi^t\rangle$ is in the biorthonormal decomposition form, and the biorthonormal decomposition theorem guarantees that this state is unique. Hence:

$$|\Psi^t\rangle = \hat{U}_1 \otimes \hat{U}_2 \left(\sum_i c_i \psi_i \otimes \eta_i \right),$$

and:

$$\hat{U}_{1+2} = \hat{U}_1 \otimes \hat{U}_2.$$

This lemma can be extended into a full theorem if the coefficients c_i are time-independent. It is possible then to prove that if $\hat{U}_{1+2} = \hat{U}_1 \otimes \hat{U}_2$ then $W_1 = \hat{U}_1 W_1 \hat{U}_1^{-1}$ and $W_2 = \hat{U}_2 W_2 \hat{U}_2^{-1}$. A full proof of this theorem, with extensions to mixed states of the composite, is contained in [106].

Chapter 4

Quantum Causation

4.1 Quantum Correlation Phenomena

In the previous chapter I provided one illustration of scientific-theory application. In this chapter I explicitly compare methodological strategies employed in application and confirmation. In doing so, I hope to be further strengthening the case for a sharp methodological, as well as epistemic, distinction between application and confirmation.

In the first section I review the correlations between distant quantum particles described by Einstein, Podolsky and Rosen in their famous 1935 'EPR' paper[1]. In the following two sections I am concerned with the philosophical debate concerning possible explanations of these correlations. Specifically I ask the question: can the EPR correlations be explained by means of causal models? Two kinds of models are possible: direct-cause models and common-cause models. I address an argument due to Van Fraassen against common cause models of the EPR correlations; and I concentrate on an objection to direct-cause models due to Arthur Fine. I argue that while Van Fraassen is concerned with the empirical adequacy of causal theories in general, Fine's argument relates to the application of quantum theory. It is

[1]Einstein, Podolsky and Rosen, '*Can Quantum Mechanical Description of the World Be Considered Complete*' [52].

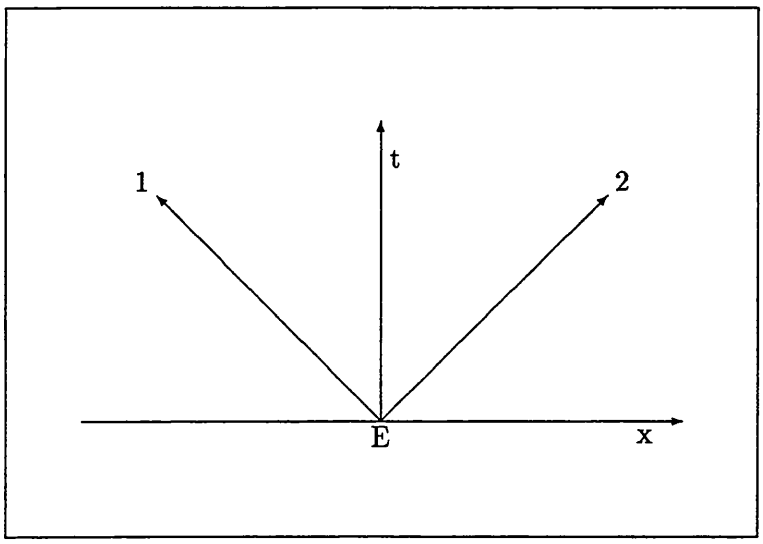


Figure 4.1: EPR particles in Minkowski Spacetime

not surprising then, that the responses to these arguments differ methodologically. Strategies successfully employed in restoring empirical adequacy may have no use in application.

4.1.1 The Einstein-Podolsky-Rosen Correlations

The version of the EPR experiment that usually gets discussed in the literature is due to David Bohm[2]. In Bohm's version, two particles (an electron-positron pair perhaps) are created at a source in a decaying event E. The particles ("1", "2") travel in opposite directions. In a Minkowski spacetime diagram they travel along paths symmetric to the time axis (see figure 4.1). The initial angular momentum, before the creation event, is zero. After the creation event, the spin of both particles must add up to zero: the spin values for the particles must be correlated. A particle's spin can be measured

[2]Bohm [15, pp.614-623].

by means of a Stern-Gerlach apparatus. This is a magnetometer that impresses a force on the particle proportional to its spin value, correlating the particle's position to its spin value at the time it interacts with the apparatus. The magnetometer can be rotated and set along any direction in space; different orientations of the magnetometer will result in measurements of different components of the particle's spin.

The particles' spin values are measured on both wings, and the measurement events are spacelike separated. According to quantum mechanics there are two possible values of the spin component along any one direction θ: spin-up or spin-down. (I refer to these as $\uparrow_\theta$ and $\downarrow_\theta$ respectively). Quantum mechanics describes the state of the composite system of both particles by means of the so-called singlet state:

$$\phi = \frac{1}{\sqrt{2}}|\uparrow_{1\theta}\rangle|\downarrow_{2\theta}\rangle - \frac{1}{\sqrt{2}}|\downarrow_{1\theta}\rangle|\uparrow_{2\theta}\rangle \qquad (4.1.1)$$

where the subscripts refer to particles "1", "2". The theory gives two types of probabilistic predictions. First, it gives predictions regarding the outcomes of measurements made on either particle. The quantum mechanical statistical algorithm[3] can be employed to find out the probabilities for possible outcomes of measurements on either particle "1" or particle "2". For instance, the probability of finding spin-up on a measurement of either particle is:

$$\begin{aligned} prob(\uparrow_{1\theta}) &= <\phi|\uparrow_{1\theta}\rangle\langle\uparrow_{1\theta}|\phi> = \left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2} \\ &= <\phi|\uparrow_{2\theta'}\rangle\langle\uparrow_{2\theta'}|\phi> = prob(\uparrow_{2\theta'}). \end{aligned} \qquad (4.1.2)$$

Second, quantum mechanics also gives, by means of the so-called Lüders' rule, probabilities for state transitions in any one system. For a system in state W on a Hilbert space H, if H_a, H_b are subspaces of H, then Lüders' rule asserts that there exists a probability measure P such that:

$$P(H_a/H_b) = \frac{Tr(P_B W P_B P_A)}{Tr(W P_B)}$$

[3]See chapter 3, appendix 1.

where P_A, P_B are projectors upon subspaces H_a, H_b, defined uniquely in virtue of the isomorphism between projectors and subspaces of Hilbert space. I shall not explain Lüders' rule in detail; I concentrate only on its consequences for conditional probabilities in the singlet state. In the case of composite systems, Lüders' rule reduces to the fourth axiom of the probability calculus for conditional probability[4]:

$$prob\,(S_1/S_2) = \frac{prob\,(S_1 \& S_2)}{prob\,(S_2)} \tag{4.1.3}$$

Lüders' rule gives, in this case, *conditional* probabilities establishing correlations between outcomes of measurements on both particles. It confirms that the spin values of the particles are correlated. Suppose that I measure spin along the θ direction on the first particle, S_1, and immediately after I measure spin along the θ' direction on the second particle, S_2. And suppose, moreover, that the outcome of the first measurement is spin-up ($\uparrow_{1\theta}$). What is then the probability of obtaining spin-up in a measurement on the second particle, S_2? The quantum mechanical algorithm gives the following joint probability:

$$prob\,(\uparrow_{2\theta'} \,\&\, \uparrow_{1\theta}) = \frac{1}{2} sin^2 \frac{1}{2} \widehat{\theta\theta'}$$

By employing Lüders' rule together with equation 4.1.2, we can then find the following expression for the conditional probability:

$$prob\,(\uparrow_{2\theta'} \,/\, \uparrow_{1\theta}) = \frac{prob\,(\uparrow_{2\theta'} \,\&\, \uparrow_{1\theta})}{prob\,(\uparrow_{1\theta})} = sin^2 \frac{1}{2} \widehat{\theta\theta'} \tag{4.1.4}$$

which yields the EPR correlations for most values of θ, θ':

$$prob\,(\uparrow_{2\theta'} \,/\, \uparrow_{1\theta}) \neq prob\,(\uparrow_{2\theta'})$$

In the special case where $\theta = \theta'$, we obtain:

$$prob\,(\uparrow_{2\theta} \,/\, \uparrow_{1\theta}) = \frac{1}{2} sin^2 \frac{1}{2} \widehat{\theta\theta} = 0.$$

[4]See for instance Hughes [89, pages 349-350].

And correspondingly,

$$prob(\downarrow_{2\theta} / \uparrow_{1\theta}) = 1 - prob(\uparrow_{2\theta} / \uparrow_{1\theta}) = 1.$$

Hence, we can see that whenever spin is measured along the same direction on both particles, the singlet state has got built into it a perfect anticorrelation between their spin values. If we measure spin along the θ direction on the first particle, and find that the outcome corresponds to spin-up, we are in a position to predict with certainty that the outcome of a subsequent measurement of spin along θ on the second particle will be spin-down.

4.1.2 Bell's Result

The simplest explanation of these distant correlations would be to assume, as in any analogous case of correlation in classical physics, that the particles already possess definite values for spin as they leave the source. John Bell famously showed that this explanation is not available for the quantum correlations. Bell[5] derived mathematical relations between the joint expectations of values of spin measured along different directions. These relations (Bell's inequalities) would have to be satisfied by any model of the EPR correlations that assumed (1) that the particles have well-defined values of spin at the time they interact with the apparatus, and (2) that on neither wing does the outcome of the measurement event depend on any spacelike separated events. A model that satisfies (1) and (2) is often called a *local realistic* model. It is supposed to be *realistic* because it satisfies requirement (1): measurements may reveal spin values, but cannot generate them[6]. And it is supposed to be *local* because it satisfies requirement (2): the outcomes of

[5]Bell [12].

[6]Whether requirement (1) is actually essential for a *realist* construal of a scientific theory is uncertain. The requirement is only vaguely connected, if connected at all, to the kinds of realism about scientific theory that I have so far discussed in the Thesis. A more accurate and neutral terminology would be 'local and *value-definite* models'. However, for the sake of consistency with other writers, I will continue to use the standard terminology.

measurement do not depend on any spacelike separated events. This locality requirement has often been expressed as the requirement that a certain factorizability condition, which I discuss below, be satisfied.

There are essentially two kinds of local realistic models: deterministic and stochastic. The deterministic model is the simplest one described above: as they leave the source, the particles already possess definite values of spin. The possession of these values is not described by the singlet state, so these values must be fixed by some 'hidden' state of the system, usually represented by λ. Thus, in an ensemble of systems prepared in the same singlet state, the correlations will be determined by the distribution of λ:

$$P(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'}) = \int A(\theta, \lambda) B(\theta', \lambda) \rho(\lambda) d\lambda$$

where $A(\theta, \lambda)$, $B(\theta', \lambda)$ are the corresponding characteristic functions (for instance: $A(\theta, \lambda) = 1$ if and only if a measurement of spin along direction θ made on a system in 'hidden' state λ yields outcome $\uparrow$, and $= 0$ iff it does not yield such an outcome, i.e. if it yields $\downarrow$ instead.)

In a stochastic local model, by contrast, the hidden state, λ, does not fix the values of spin of each particle as they are ejected from the source. Instead λ fixes probability distributions for the different values. First, it fixes single probabilities for values of spin of particle "1" or particle "2": $Prob_\lambda(\uparrow_{1\theta})$, $Prob_\lambda(\uparrow_{2\theta'})$, $Prob_\lambda(\downarrow_{1\theta})$, $Prob_\lambda(\downarrow_{2\theta'})$, etc. Second, λ also fixes joint probability distributions for both particles: $Prob_\lambda(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'})$, $Prob_\lambda(\downarrow_{1\theta} \,\&\, \uparrow_{2\theta'})$, etc. Hence, although the 'hidden' state does not determine the precise outcome, it does determine the overall joint probability distribution for outcomes of measurements on both wings:

$$P(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'}) = \int Prob_\lambda(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'}) \rho(\lambda) d\lambda.$$

Bell describes a particular sense in which a model like this could be described as *local*: the distribution over λ is assumed to be independent of the settings of either measurement apparatus. This is to ensure that '*the result B for particle 2 does not depend on the setting θ of the magnet for particle 1, nor A on θ'*' (Bell [12, page 404]). For, in Aspect's 1982 experiment[7], which

[7] Aspect et al [5].

confirmed the quantum mechanical predictions, the setting on one wing and the outcome event in the distant wing are spacelike separated. Hence, in the context of Aspect's experiment, the *locality* assumption (requirement (2)) entails that the outcome of a measurement on either wing cannot depend upon the setting of the distant apparatus. This locality assumption[8] has traditionally been cast in the form of a condition of stochastic independence:

$$Prob_\lambda(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'}) = Prob_\lambda(\uparrow_{1\theta}) \times Prob_\lambda(\uparrow_{2\theta'}). \qquad (4.1.5)$$

Hence, as a result, the overall joint probability distribution factors out:

$$P(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'}) = \int Prob_\lambda(\uparrow_{1\theta} \,\&\, \uparrow_{2\theta'})\rho(\lambda)d\lambda =$$
$$= \int Prob_\lambda(\uparrow_{1\theta}) \times Prob_\lambda(\uparrow_{2\theta'})\rho(\lambda)d\lambda.$$

Bell showed that either kind of model ('deterministic' or 'stochastic factorizable') is committed to the Bell inequalities which are violated by experiment, and equally violated by the quantum mechanical predictions. It would appear that no model that begins by assuming that each particle has a definite value of spin at the time of measurement, –one which does not depend on the setting of the distant measurement apparatus–, can account for the EPR correlations. It has been claimed that because of this the violation of the Bell inequalities by the EPR correlations refute '*local realism*'. This conclusion is, however, far from being universally accepted. Arthur Fine in particular, has contested the assumption that the condition of stochastic independence (condition 4.1.5) captures the notion of physical locality; he has constructed models of the EPR correlations that purport to be local, yet are not factorizable[9] . My aim, however, is not to discuss locality, but causality. Hence I shall not be concerned with the Bell inequalities themselves.

[8] As a matter of fact, three different and independent conditions need to be assumed to derive Bell's result. I discuss these conditions in the following section, and again in section 4.2.2.

[9] Fine [61].

Rather I shall be concerned with possible causal readings of the condition that Bell notoriously assumed in the derivation of the Bell inequalities. This condition, expressed in equation 4.1.5, is nowadays known as *factorizability.*

4.1.3 Factorizability

Suppose that s_1 and s_2 are the outcomes of the measurements on the first and the second particle. We know that these outcomes will exhibit statistical correlation: $P(s_2/s_1) \neq P(s_2)$, or in other words, that $P(s_2\&s_1) \neq P(s_2) \times P(s_1)$. Suppose also that a is the setting of the apparatus that makes measurements on the first particle, and b the setting of the apparatus that makes measurements on the second particle. These are settings of the apparatus to measure spin along different directions θ, θ', etc. The factorizability condition, required to derive Bell's inequalities, is then the following condition (in the larger probability space):

$$prob\,(s_1\&s_2/a\&b\&\lambda) = prob\,(s_1/a\&\lambda) \times prob\,(s_2/b\&\lambda) \qquad (4.1.6)$$

The condition says that when we conditionalise on the hidden state and both settings the probabilities for outcomes on both wings of the experiment, conditional on the hidden state and the corresponding setting, factor out; the original correlation disappears, as s_1 and s_2 become statistically independent of each other.

Why would factorizability be a natural condition to impose on local realistic models? Abner Shimony [121] has argued, following Jon Jarrett [90], that factorizability is logically equivalent to the conjunction of two other probabilistic conditions, *Parameter Independence* (PI) and *Outcome Independence* (OI)[10]. (PI) asserts that the probabilities for outcomes of measurements with a fixed setting on one wing are independent of the setting

[10]Shimony's conditions, which I adopt here, are somewhat different from Jarrett's. Jarrett introduces variables α and β to describe any further features of the measurement apparati that may affect the outcomes. He then conditionalises on all of a, b, α, β and λ. Jarrett's *locality* condition is then analogous to parameter independence: $P(s_1/a\&b\&\alpha\&\beta\&\lambda) = P(s_1/a\&b'\&\alpha\&\beta'\&\lambda)$, while his *completeness* condition corresponds to outcome independence: $P(s_1/s_2\&a\&b\&\alpha\&\beta\&\lambda) = P(s_1/a\&b\&\alpha\&\beta\&\lambda)$. Shi-

of the distant apparatus, if we keep λ fixed:

$$prob(s_1/a\&b\&\lambda) = prob(s_1/a\&\lambda)$$
$$prob(s_2/a\&b\&\lambda) = prob(s_2/b\&\lambda) \qquad (4.1.7)$$

While (OI) asserts that the outcomes on one wing are statistically independent of the outcomes in the other wing, if we keep the hidden state and both settings fixed:

$$prob(s_1/s_2\&a\&b\&\lambda) = prob(s_1/a\&b\&\lambda)$$
$$prob(s_2/s_1\&a\&b\&\lambda) = prob(s_2/a\&b\&\lambda) \qquad (4.1.8)$$

I shall return to the interpretation and implications of these conditions in section 3. For now, I just want to point out that factorizability can be derived from (OI) and (PI), given the axioms of the probability calculus, as follows:

$$prob(s_1\&s_2/a\&b\&\lambda)$$
$$= prob(s_1/a\&b\&\lambda\&s_2) \times prob(s_2/a\&b\&\lambda) \qquad (4.1.9)$$
$$= prob(s_1/a\&b\&\lambda) \times prob(s_2/a\&b\&\lambda) \qquad (4.1.10)$$
$$= prob(s_1/a\&\lambda) \times prob(s_2/b\&\lambda) \qquad (4.1.11)$$

Expression 4.1.9 follows from equation (4.1.3), which is the definition of conditional probability. Expression 4.1.10 can be derived from 4.1.9 by (OI); while expression 4.1.11 can be derived from expression 4.1.10 by (PI).

Hence factorizability is the conjunction of (i) outcome independence, (ii) parameter independence, and (iii) the calculus of probabilities, including the definition of conditional probability. Factorizability enables us to derive the Bell inequalities, and the Bell inequalities are violated by experiment. But by the derivation above, a violation of factorizability is equivalent to a violation of either (i) outcome independence or (ii) parameter independence

mony's conditions can be made equivalent to Jarrett's by taking a, b to represent both the setting and further relevant factors of each apparatus.

or (iii) the calculus of probability, or (iv) some combination of (i), (ii), (iii). Although the definition of conditional probability fails in general in the standard axiomatisations of quantum probability, its application in this context is justified because we are dealing with experimental statistics, not quantum probabilities. Thus, any empirically adequate theory must reject either (OI) or (PI), or both.

It is not difficult to show that quantum mechanics satisfies (PI). This follows from the spherical symmetry of the singlet state:

$$\begin{aligned}\phi &= \frac{1}{\sqrt{2}}|\uparrow_{1\theta}\rangle|\downarrow_{2\theta}\rangle - \frac{1}{\sqrt{2}}|\downarrow_{1\theta}\rangle|\uparrow_{2\theta}\rangle \\ &= \frac{1}{\sqrt{2}}|\uparrow_{1\theta'}\rangle|\downarrow_{2\theta'}\rangle - \frac{1}{\sqrt{2}}|\downarrow_{1\theta'}\rangle|\uparrow_{2\theta'}\rangle\end{aligned}$$

for any θ, θ'. The first measurement, whether made on particle "1" or "2", gives outcome $\uparrow$ or $\downarrow$ with probability $p = 1/2$, whatever direction spin is measured along on either particle. The probability for an outcome of a measurement of spin in one wing of the experiment is independent of the setting of the distant apparatus: (PI) holds. And hence, as either (PI) or (OI) must fail, in quantum mechanics (OI) fails. The suggestion has been made, by Shimony, Ballentine and Jarrett, that we should follow the example of quantum mechanics and reject the condition of outcome independence altogether. A number of arguments have been given for this rejection; I review some of them in section 4.3. But now, I turn to the central topic of this chapter: can we give a *causal* account of these experimental correlations? Do the EPR correlations admit causal models? To answer this question we must look into empiricist theories of causation, and determine if the conditions that those theories lay out are met in the EPR case.

4.2 The Principle of the Common Cause

The most influential empiricist theory of causality[11] is due to Hans Reichenbach, who proposed a concept of causation as a formal feature of probabilis-

[11] By an *empiricist* theory I mean, roughly, any theory that aims to explicate the notion of causation by broadly complying with Humean strictures, i.e. by eschewing any explicit

tic models. Reichenbach's conception was inspired as much by Rudolph Carnap's distinction between the formal and material mode of speech[12], as by the Humean critique of necessary connection. The objects of the material mode are concrete entities in the physical world, and their properties. By contrast, the objects in the formal mode are not entities in the physical world; they are instead words, sentences, logical symbols, syntactic operations –they are linguistic entities.

According to Carnap, the correct mode of speech in philosophy is the formal mode. In the formal mode there are no unsolvable problems. If a so-called philosophical problem, generated in the material mode, has no translation into the formal mode then either it is a scientific problem in disguise, and not a philosophical one, or it is not a genuine problem at all. For example, the dispute over the existence of numbers is, according to Carnap, a pseudo-dispute. The realist, such as Peano or Hilbert, asserts that numbers are primitive objects. But for Russell and Whitehead numbers are classes of classes. When these two contradictory assertions are translated into the formal mode, they cease to be contradictory: each one of them is true in a different language, or syntactic frame. In Principia numerical expressions are class expressions of the second-order; in Peano's system numerical expressions are not second-order expressions, but irreducible elementary expressions[13].

---

appeal to necessary regularities or counterfactuals. There is a more recent trend to refer to a theory of causation as empiricist if its account of causation is dictated by science, not metaphysics –in this respect, the debate between Philip Dowe [48] and Wesley Salmon [119] is instructive.

[12]This point is argued by Cartwright, see [27, 158-170]. Carnap introduced the distinction between the formal and the material modes in his 1935 *Logische Syntax Der Sprache* [23].

[13]This is Carnap's own example in [24, pp. 76-77].

4.2.1 Reichenbach's Formal Conditions

The suggestion, then, was that long-lasting philosophical debates concerning causation were similarly underpinned by disputes concerning the correct expression of the fundamental assertions of the theory. An analogous purgative exercise could then be carried out on causation, by divesting this notion of its usual formulations in terms of powers, or dispositions, and giving an appropriate definition in the formal mode of speech. In his 1956 book, 'The Direction of Time' [117], Hans Reichenbach undertook this task. Reichenbach proposed a number of features that a probabilistic model would need to possess in order to qualify as a causal model. On his view, the concept of causation could be entirely captured by describing some formal features of our representations of empirical regularities. Reichenbach proposed a theory of causation that would apply to statistical representations as well as deterministic ones. The aim was to define formally the causal structure underlying every correlation. Reichenbach's central notion was the *Principle of the Common Cause*:

> If an improbable coincidence has occurred, there must exist a common cause. (Reichenbach [117, page 157].)

Regarding this principle, two remarks are in order. First, by '*improbable*' Reichenbach does not just mean a coincidence with low prior probability. He has in mind a coincidence of two events of types A and B, such that there is no reason to suspect that A directly causes B or B directly causes A, i.e. there is no reason to suspect a direct causal link. The second remark is that, according to Reichenbach, it is not the case that a coincidence is genuine if and only if it is lawlike, in some metaphysically robust sense. Rather, a coincidence is genuine if it is an instance of some empirically established statistical correlation between event-types. Although two event-tokens may be coincident, they cannot, of course, exhibit correlation; correlations can hold among event-types only. To establish a correlation it is necessary to determine, by inductive means, and on the basis of observed regularities, a probability distribution $P(A)$ over events of kind A, a probability distribu-

tion $P(B)$ over events of kind B, and a joint distribution $P(A\&B)$ for the joint occurrence of A's and B's. Events of types A and B are said to be correlated if and only if:

$$P(A\&B) > P(A) \times P(B) \tag{4.2.12}$$

This condition is, barring the case of probability zero, provably equivalent to the following two conditions on the conditional probabilities: $P(B/A) > P(B)$, and $P(A/B) > P(A)$, barring the case of probability zero. It is clear that a correlation makes a token of A more likely to occur if a token of B has occurred, and viceversa.

The Principle of the Common Cause then asserts that any coincidence between two tokens of A and B must be explainable by either showing a direct causal connection between the types A and B, –so that the coincidence ceases to be 'improbable'–, or by pointing to a common cause C of both event types. That is, coincidences are to be explained by embedding them in observed correlations between variables in some causal structure. The main problem with this mode of explanation concerns causal inference. For any set of correlations always underdetermine its underlying causal structure. The existence of a correlation does not give grounds to assert the existence of a direct-causal link between the correlated events, *for it is always possible to postulate further structure.* Thus, a correlation between present smoking ('*smoking now*') and lung cancer ('*cancer now*') does not entail that 'smoking now' causes 'lung cancer now', for there is a hidden factor, namely a history of smoking, that is a common cause. But, then, for the same reason, no correlation ever gives grounds to infer any particular common cause. For there could be still further, underlying common causes. A patient's smoking history, his present lung cancer, and his present smoking may all be caused by a secret gene. (Figure 4.2 serves to illustrate this example. For if A is present smoking, B is present lung cancer, and D is the smoking history of the patient, it is still possible that D is not the cause of either A or B, but that they are all caused by C, the secret gene).

To reveal the genuine causal structure in probabilistic models one has to identify the underlying common causes; if we want to remain empiricists

we need to be able to identify the common causes *from the statistics alone.* The issue of identification turned out to be an exceedingly difficult one. Reichenbach's ingenious idea was that common causes must *screen off* their effects from each other. Reichenbach's definition[14] is as follows: *C screens off A from B iff* $P(A\&B/C) = P(A/C) \times P(B/C)$. Put informally, the condition says that the statistical correlation between the effects becomes irrelevant to the assessment of the occurrence of either effect, when the common cause is taken into account. Conditionalisation on the common cause should *wash out* the original correlation between its effects. In perhaps more conspicuous notation we can write the following condition (note that this condition is symmetric in A and B):

Definition 8 (Screening Off) *C screens off A from B iff* $P(A/B\&C) = P(A/C)$.

This condition is provably equivalent to Reichenbach's own[15], and I will adopt it instead. Screening off is a purely probabilistic condition: no causal notions are involved. But Reichenbach thought that screening off reveals information about causal structure. His intuition was that common effects always occur in *closed conjunctive forks*, while common causes may occur in *open conjunctive forks.* A closed fork is a parallelogram with causally related variables in each vertex; so it requires four variables. Reichenbach defined an open conjunctive fork as a set of three variables A, B C that satisfy three conditions. First, A and B exhibit correlation (condition 4.2.12); second, C screens A from B (definition 8); and third, C temporally precedes both A and B[16]. In such case, C is the common cause of A and B. I will focus on

[14]The definition is given in [117, page 159], and the terminology of screening off introduced in [117, page 189].

[15]The proof of the equivalence is very simple: If $P(A\&B/C) = P(A/C) \times P(B/C)$, then $P(A\&B/C) = P(A\&B\&C)/P(C) = P(A/B\&C)P(B/C) = P(A/C) \times P(B/C)$. Hence $P(A/B\&C) = P(A/C)$, and similarly for $P(B/A\&C) = P(B/C)$. The converse implication is trivial.

[16]As a matter of fact, the temporal precedence of causes was a consequence of Reichenbach's theory of time. Reichenbach defined the direction of time in terms of the

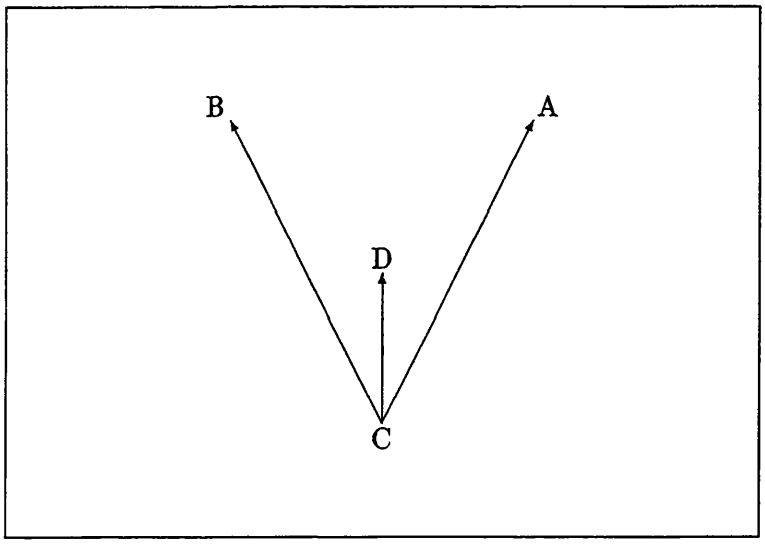


Figure 4.2: Screening-off is not a Sufficient Condition

the screening off condition. Is this a sufficient condition on common causes, a necessary condition, or both? In the causal structure pictured in figure 4.2, variable D is not a common cause of A and B although it satisfies all the criteria for a common cause (A and B are correlated, D precedes both, and it screens off A from B[17]). The true common cause is C, which is a cause of all A, B, and D.

asymmetries of causation, essentially by stipulating that time always flows in the direction of open conjunctive forks –and therefore ruling out backwards-in-time causation by definition. The reduction of time to causation is a complex and controversial issue; as I do not wish to prejudice the issue, I take the temporal precedence of C over both A and B to be an independent condition on common causes. This condition is indeed waived in some theories of probabilistic causation, most notably in Patrick Suppes's [127].

[17]This is only true if C screens off too. For if C did not screen off A from B then neither would D. But we are here *assuming* that C is the true common cause, and on Reichenbach's theory, as we shall shortly see, every common cause screens off.

It becomes clear that Reichenbach cannot really have meant that we can infer to the existence of a common cause from the statistics alone. An assumption of completeness is required to make screening off into a sufficient condition on common causes. If we knew that A, B and C are the only causally relevant variables then, according to Reichenbach's account we would also know that, if they form a conjunctive fork, C is the common cause of A and B. But the assumption of completeness is unwarranted, for at least two reasons. First, complete knowledge of the full set of causal factors is rarely, if ever at all, available in practice. Second, the completeness assumption would have gone against the form of empiricism embraced by Reichenbach: in order to arrive at knowledge of causes through statistics one would need to know in advance the full causal structure. For both these reasons Reichenbach settled for necessity rather than sufficiency. Although it is not true that every screener-off is a common cause, in Reichenbach's theory every common cause must be a screener-off. The discovery of causal structure was to proceed negatively, as it were, on the basis of violations of the screening-off condition.

4.2.2 Van Fraassen against Causal Realism

Bas Van Fraassen[18] adopted Reichenbach's conditions in order to advance an argument against the thesis of *causal realism* –the thesis that there are causal mechanisms underlying all empirical lawlike regularities. Causal realism is false, according to Van Fraassen, because no causal theory can ever be empirically adequate. There are some correlations that no causal theory could ever account for. Specifically, the EPR correlations cannot be embedded in any statistical model that complies with Reichenbach's formal conditions. It then follows, taking embedding as a *necessary* as well as sufficient condition for empirical adequacy, that no causal model can account for (i.e. be empirically adequate of) the EPR correlations.

[18]Van Fraassen [136].

Van Fraassen's argument has convinced many, including some firm advocates of causal realism. Wesley Salmon, for instance, writes:

> Van Fraassen argues cogently, on the basis of Bell's inequalities and relevant experimental results, that 'there are well–attested phenomena which cannot be embedded in any common-cause model'. [...] When we ask for the causal mechanism involved in the production of the EPR correlations, we find ourselves at a loss. (Salmon, [118, pp. 251, 254])

In the EPR experiment, measurement outcomes are correlated. It is possible to interpret hidden state theories as postulating a common cause λ, the creation event at the source, that gives rise to the correlations. Van Fraassen writes down three conditions that any putative common cause in EPR must satisfy. Suppose that the common cause is ψ, the quantum state of the two particles at the source –in the EPR case, this is the singlet state[19]. Van Fraassen first two conditions are a precedent of (OI) and (PI) respectively. The first condition, –which Van Fraassen names '*causality*'– is inspired by Reichenbach's screening off condition:

Definition 9 (Van Fraassen's 'Causality')

$$prob\,(s_1/s_2\&a\&b\&\psi) = prob\,(s_1/a\&b\&\psi)$$
$$prob\,(s_2/s_1\&a\&b\&\psi) = prob\,(s_2/a\&b\&\psi)$$

The second condition is named '*hidden locality*':

Definition 10 (Van Fraassen's 'Hidden Locality')

$$prob\,(s_1/a\&b\&\psi) = prob\,(s_1/a\&\psi)$$
$$prob\,(s_2/a\&b\&\psi) = prob\,(s_2/b\&\psi)$$

[19]This is a substantial assumption, as the state could also evolve stochastically. Jarrett has argued that the Bell inequalities could be violated due to indeterministic time-evolution (see Jarrett [91]). In general one may prefer to use the state of the particles at the time of measurement, instead of the state at the source as I do here.

The third condition is designed to rule out any possible dependency of the common cause on either of the settings (recall that in the derivation of the Bell inequalities, it is assumed that the probability distribution over the hidden variable, $\rho(\lambda)$, does not depend on any of the settings):

Definition 11 (Van Fraassen's 'Hidden Autonomy')

$$prob\,(\psi/a\&b) = prob\,(\psi)$$

Van Fraassen first claims that a violation of either 'hidden locality' or 'hidden autonomy' would be incompatible with a common cause explanation of the correlations:

> If the probability of a given outcome at [one wing] is dependent not merely on the putative common cause, but also on what happens at [the other wing], or if the character of that putative common cause itself depends on which experimental arrangement is chosen (even if after the source has been constructed) then I say that the two outcome-events have not been traced back to a common cause which explains their correlation. (Van Fraassen, op.cit [136, page 105]).

On the other hand, 'causality' is required by Reichenbach's principle of the common cause. This is easiest to see in the simplest case where the apparatus settings are fixed in the same orientation in both wings of the experiment. 'Causality' is then nothing but the demand that the common cause screens off:

$$P\,(s_1/s_2\&\psi) = P\,(s_1/\psi)$$
$$P\,(s_2/s_1\&\psi) = P\,(s_2/\psi)$$

The conjunction of 'causality', 'hidden locality', and 'hidden autonomy' is committed to the Bell inequalities, which are violated by experiment; hence at least one conjunct must be false. In particular, if the quantum state itself is to represent the total state of the common cause, we get a

straightforward violation of the 'causality' condition by simply considering the EPR correlations. For the singlet state cannot screen off the correlations between the outcomes. Here we can take a smaller probability space, writing the quantum state as a subscript to the probability function, and hence avoiding having to assign probabilities to impossible propositions that state that the combined particle-pair system after measurement is both in state ψ and in whatever state corresponds to outcomes s_1 or s_2, and also avoiding having to assign probabilities for the occurrences of quantum states. We then make the natural analogue of the previous conditions; and find:

$$P_\psi(s_1/s_2) \neq P_\psi(s_1)$$
$$P_\psi(s_2/s_1) \neq P_\psi(s_2)$$

Hence, trivially, in the EPR experiments, the quantum state of the particle-pair cannot be the common cause. And, more generally: if 'causality', 'hidden locality' and 'hidden autonomy' are necessary conditions on a common cause, it is impossible to embed the EPR correlations into a causal model of the sort described by Reichenbach, –regardless of whether the common cause is the quantum state, or a more complete hidden state. To conclude: by assuming only Reichenbach's conditions on common causes, Van Fraassen derives a striking refutation of causal realism by the EPR correlations.

4.2.3 Causation in a Probabilistic World

It is now time to look more carefully into Van Fraassen's presuppositions. In particular, is screening-off really necessary for a common cause? In a series of recent papers, Nancy Cartwright[20] has shown that screening off is too restrictive a condition on common causes. Screening off is a necessary condition on *deterministic* common causes only. In the most general case,

[20]Cartwright [28], [32], [33], and particularly [27, section 3.3]. Cartwright's ideas have been explicitly applied to the EPR correlations in a joint paper with Hasok Chang [36], which also deals with many of the issues that I discuss in this chapter.

when a cause acts probabilistically to generate its effects, screening off is obeyed by just a very special subset of all possible common causes.

In order to bring this point home, let me first consider a remarkable consequence of Van Fraassen's analysis of the EPR correlations: when applied to the perfect correlations in the singlet state, screening off entails determinism[21]. We begin by rewriting the EPR perfect anti-correlations as:

$$P(s_1 = \uparrow \,\&\, s_2 = \uparrow) = 0 = P(s_1 = \downarrow \,\&\, s_2 = \downarrow).$$
$$P(s_1 = \uparrow \,\&\, s_2 = \downarrow) = 1 = P(s_1 = \downarrow \,\&\, s_2 = \uparrow).$$

When we have probabilities one or zero, we can conditionalise on anything that has non-zero probability. So we can conditionalise on λ, which we assume to have non-zero probability, and obtain:

$$P(s_1 = \uparrow \,\&\, s_2 = \uparrow / \lambda) = 0 = P(s_1 = \downarrow \,\&\, s_2 = \downarrow / \lambda).$$
$$P(s_1 = \uparrow \,\&\, s_2 = \downarrow / \lambda) = 1 = P(s_1 = \downarrow \,\&\, s_2 = \uparrow / \lambda).$$

We now impose Reichenbach's screening off condition:

$$P(s_1 = \uparrow \,\&\, s_2 = \uparrow / \lambda) = P(s_1 = \uparrow / \lambda) \times P(s_2 = \uparrow / \lambda) = 0.$$
$$P(s_1 = \uparrow \,\&\, s_2 = \downarrow / \lambda) = P(s_1 = \uparrow / \lambda) \times P(s_2 = \downarrow / \lambda) = 1.$$
$$P(s_1 = \downarrow \,\&\, s_2 = \uparrow / \lambda) = P(s_1 = \downarrow / \lambda) \times P(s_2 = \uparrow / \lambda) = 1.$$
$$P(s_1 = \downarrow \,\&\, s_2 = \downarrow / \lambda) = P(s_1 = \downarrow / \lambda) \times P(s_2 = \downarrow / \lambda) = 0.$$

In order to satisfy all these equalities at once, either $P(s_1 = \uparrow / \lambda) = 0$ and $P(s_2 = \uparrow / \lambda) = 1$; or, alternatively, $P(s_1 = \uparrow / \lambda) = 1$ and $P(s_2 = \uparrow / \lambda) = 0$.

[21] This result can be seen an instance of Arthur Fine's theorems on factorizable stochastic models. Fine [60] has shown, following Suppes and Zanotti [129], that factorizable stochastic and deterministic hidden variable models of the quantum correlations are mathematically equivalent, in the following sense: if some set of correlations admits a factorizable stochastic model, it admits a deterministic model as well. It is clear that Fine's result must be applicable to Reichenbach's screening-off condition; for recall that Reichenbach originally expressed his requirement as the condition that the outcomes *factor out* conditional on the common cause.

It follows, in either case, that in order to satisfy 'causality' λ must give probability 1 or 0 to *all* possible outcomes in both wings of the experiment. That is, λ must function as a deterministic common-cause: it is hardly surprising that Bell's inequalities are violated!

Reichenbach's analysis of causation, which was originally intended to apply to probabilistic and deterministic causes alike, is actually committed to determinism in this particular context. Any putative common cause that –as demanded by Van Fraassen–, obeyed Reichenbach's screening off condition would have to act deterministically in each wing of the experiment to generate the EPR correlations. However, there is nothing in the concept of probabilistic causation to indicate that a common cause must determine its effects in this manner.

Consider a simple case of purely probabilistic causation: a factor z causes a factor x with, say, 75% probability. In a representative sample the presence of z is sufficient for x in 75% of the cases, and fails to produce x 25% of the times. We can represent the action of z to produce x by means of the symbol $\hat{a}_{zx}$, an indicator function that takes value 1 whenever z fires to produce x and value 0 whenever z fails to fire to produce x. We can then fix the expectation of the action of z to produce x accordingly: $Exp(\hat{a}_{zx}) = P(\hat{a}_{zx}) = \frac{3}{4}$. If z is x's only cause then $P(x) = P(\hat{a}_{zx})P(z) = \frac{3}{4}P(z)$; this equation determines the frequency of occurrence of x completely.

Consider now a three-variable structure, where z is a common cause of x and y, and there is no direct causal link between the latter two. We write:

$$P(x) = P(\hat{a}_{zx})P(z)$$
$$P(y) = P(\hat{a}_{zy})P(z)$$

If a common cause is really probabilistic, its action to bring about one effect may bear any relation whatever to its action to bring about any other. For instance, it is not required, in the example above, that $P(\hat{a}_{zx}) = P(\hat{a}_{zy}) = \frac{3}{4}$. And indeed, the action of z to produce x may be totally independent of its action to produce y; in such case the joint expectation of its indicator functions factorizes: $P(\hat{a}_{xy}\&\hat{a}_{xz}) = P(\hat{a}_{xy}) \times P(\hat{a}_{xz})$. As a consequence

the joint probability of x and y, conditional on the common cause, also factorizes, and screening off is satisfied by this common cause:

$$P(x\&y/z) = \frac{P(\hat{a}_{zx}\&\hat{a}_{zy})}{P(z)} = \frac{P(\hat{a}_{zx})}{P(z)} \times \frac{P(\hat{a}_{zy})}{P(z)} = P(x/z) \times P(y/z)$$

On the other hand some stochastic dependence between the actions of the common cause cannot be ruled out: z's action to generate x may overlap with its action to generate y. In fact, there could be a perfect correlation between $\hat{a}_{zx}$ and $\hat{a}_{zy}$. A particular instance of this type of cause is one that never produces one effect without the other; a cause that invariably either acts to produce all of its effects, or fails to produce any. The joint expectation of this cause's actions won't factorize ($P(\hat{a}_{zx}\&\hat{a}_{zy}) = P(\hat{a}_{zx}) \neq P(\hat{a}_{zx}) \times P(\hat{a}_{zy})$), and the common cause won't generally screen off:

$$P(x\&y/z) = \frac{P(\hat{a}_{zx}\&\hat{a}_{zy})}{P(z)} = \frac{P(\hat{a}_{zx})}{P(z)} = P(x/z) \neq P(x/z) \times P(y/z)$$

Screening off is a necessary condition on a common cause that produces its effects independently of each other. But is not a reasonable condition to impose on a cause that exhibits some degree of correlation between its effects. Using the small probability space notation once more: the common cause z determines a probability function P_z over its effects. There is no reason why this function's distribution over the effects x and y must obey the following strict condition derivable from screening off [22]: $P_z(x\&y) = \frac{P(x\&\neg y) \times P(\neg x\&y)}{P(\neg x\&\neg y)}$. Any other combination is also possible.

The application of Reichenbach's Principle of the Common Cause to probabilistic causes requires a revision of the conditions on common causes. Nancy Cartwright has enunciated a general criterion for any common cause, whether of the deterministic, factorizable stochastic, or purely probabilistic kind[23]. To bring it closer to Reichenbach's Principle, her criterion is formu-

[22] This condition can be derived from screening off by means of the equality: $P_z(x\&y) = P_z(x) \times P_z(y) = [P_z(x\&y) + P_z(x\&\neg y)] \times [P_z(y\&x) + P_z(y\&\neg x)]$.

[23] Cartwright [28, page 187] is the earliest statement. The criterion is restated in [27, page 236], with the additional qualifications that I discuss below.

lated as a necessary and sufficient condition on the absence of a direct-link between the effects of the putative common cause. It says that x and y are not directly causally linked if and only if their joint probability is entirely due to the joint expectation of z's actions to produce them:

Criterion 4.2.1 (General Common-Cause Criterion) *There is no direct causation between x and y (i.e. $\hat{a}_{xy}$ is always zero) if and only if $P_z(x\&y) = P_z(\hat{a}_{zx}\&\hat{a}_{zy})$.*

I want to break down this criterion into its necessary and sufficient parts. I shall embrace necessity but, for the purposes of EPR, reject sufficiency. This is because the sufficient part of Cartwright's general criterion rules out a significant class of possible common cause models for the EPR correlations. Besides the necessary part is all that is really needed to show, contra Van Fraassen, that common cause models cannot be ruled out.

Criterion 4.2.2 (Necessary Condition on Common-Causes) *If z were the only cause of x, and of y, the joint probability for x and y would be equal to the joint expectation of z's actions to produce x and y: $P_z(x\&y) = P_z(\hat{a}_{zx}\&\hat{a}_{zy})$.*

Hence the necessary part of Cartwright's General Criterion asserts that if z is a common cause of x and y, *and* if there is no direct causal link between x and y, the joint probability of x and y is the joint expectation of the cause's actions. This is always true: if the common cause z is the only cause of x and y, then x and y will co-occur when and only when z fires to produce both x and y.

Criterion 4.2.3 (Sufficient Condition on Common-Causes) *If it is the case that $P_z(x\&y) = P_z(\hat{a}_{zx}\&\hat{a}_{zy})$ then neither x nor y cause one another (z is their only cause).*

The sufficient part of Cartwright's criterion is, however, not generally true. It fails in two types of cases. First, suppose that y is a direct cause of x, but that $\hat{a}_{zy} = 0$, i.e. y is never actually caused by z (see figure 4.3). Suppose

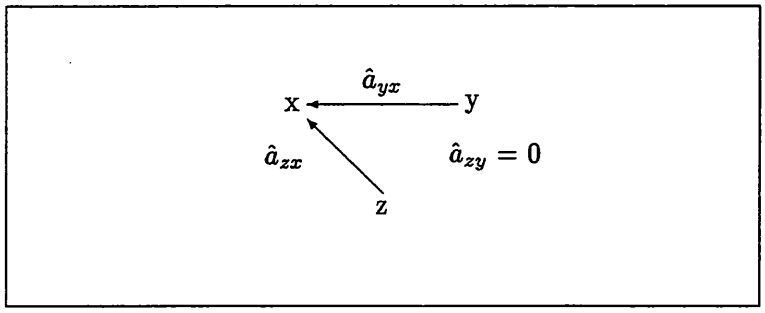


Figure 4.3: General Criterion: First Counterexample

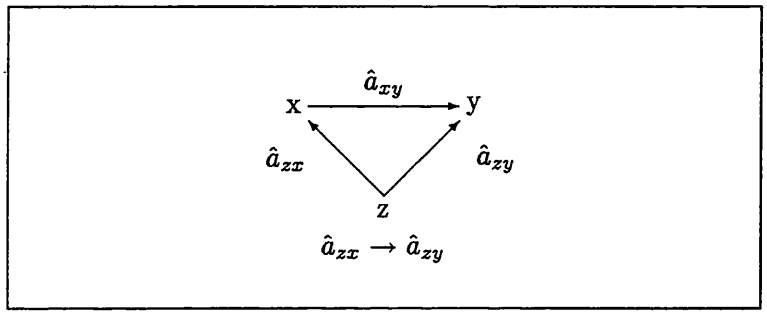


Figure 4.4: General Criterion: Second Counterexample

further that y happens never to produce x whenever z occurs, and z never to produce x whenever y occurs. (Or, alternatively suppose that z and y never occur together.) Then $P_z(x\&y) = 0 = P_z(\hat{a}_{zx}\&\hat{a}_{zy})$, and the antecedent of the criterion holds, even if the consequent doesn't.

Second, suppose the arrangement of (figure 4.4), with z as a cause of x and y, and x as a direct cause of y (possibly a deterministic one). Imagine that x, though a cause of y, is always preempted by z, i.e. $\hat{a}_{xy} \rightarrow \hat{a}_{zy}$. Suppose in addition that $\hat{a}_{zx} \rightarrow \hat{a}_{zy}$, i.e. whenever z causes x, it also causes y. Even if a direct cause link exists between x and y, the antecedent of the sufficient condition is satisfied: $P_z(x\&y) = P_z(x) = P_z(\hat{a}_{zx}) = P_z(\hat{a}_{zx}\&\hat{a}_{zy})$

In order to deal with these counterexamples, Cartwright has constrained the operation of a common cause in the following two ways[24]:

(1) $\hat{a}_{zx} \neq 0, \hat{a}_{zy} \neq 0$.

This constraint makes it clear that it is not possible to infer that z is a common cause by looking at the statistics alone. For (1) says that in order to apply the General Criterion to z, we first need to know that z is indeed a common cause. This is not a problematic constraint, though. Recall that Reichenbach's screening off was never meant as a sufficient condition on common causes –the assumption of completeness of the causal structure had to be added to make screening off sufficient. The situation is formally analogous here. Prior knowledge of the causal structure that underlies the statistics is required for Cartwright's criterion to work as a sufficient condition,

(2) $\neg(\hat{a}_{zx} \rightarrow \hat{a}_{zy}), \neg(\hat{a}_{zx} \leftarrow \hat{a}_{zy})$.

The second constraint, I think, spells more trouble. (2) rules out a common cause whose operations to cause some effect may entail some other of its operations to cause another effect. It rules out a common cause z that never acts to produce y without also acting to produce x; or viceversa, one that never acts to produce x without acting to produce y. A special case is a cause whose operations are fully coextensive, i.e. one whose operations entail each other.

We now see that Cartwright's General Criterion, as qualified, is inapplicable to causes whose actions are genuinely coextensive. This is not to claim that the criterion is inapplicable to causes that generate perfect statistical correlations, of the EPR type, between its effects. Coextensiveness of a cause's operations is, of course, not entailed by perfect correlation between its effects. A cause's effects E_1, E_2 exhibit perfect correlation iff $P(E_1/E_2)$

[24]These are essentially Cartwright's own constraints in [27, page 236], –although they are incomplete as they appear there, and require slight amendments.

= 1 or 0, but this is consistent with there being a small subset –of measure-zero– of E_1, E_2 tokens for which $E_1\&\neg E_2$, or $\neg E_2\&E_1$. By contrast, perfect coextensiveness of a cause's operations means that every single instance of E_1 is accompanied by an instance of E_2, and viceversa. So, as long as the EPR common cause's operations are not coextensive, Cartwright's qualified General Criterion can be safely applied to the EPR correlations; it is then possible to make the stronger claim that there exists no direct causal link between the wings of an EPR experiment if there is a common cause whose actions generate the EPR correlations exactly.

The point I want to make is that Cartwright's qualified General Criterion, if applied as a sufficient as well as necessary condition on a common cause, would be too restrictive: it would rule out a cause whose operations are fully coextensive –i.e. one that invariably produces its effects in tandem. I don't see any reason to rule out such a common cause explanation of the EPR correlations. Indeed, with the apparatus settings fixed in the same direction in both wings of the experiment, the simplest possible model for the EPR correlations is a common cause that in every single instance acts in both wings, producing outcomes of opposite spin value. The cause's actions may not always be reflected in the actual experimental outcomes; there may well be other interfering causes operating, and the common cause may just be a partial one. In other words, it is likely that, in any real EPR experiment the antecedent of the conditional in criterion (4.2.2) is not satisfied, so the actual experimental outcomes may fail to exhibit the perfect coextensiveness of the common cause's operations. But the point stands: it makes perfect sense, in EPR, to postulate a common cause with coextensive operations. For, *if the antecedent of criterion (4.2.2) were true the consequent would follow*, even for this perfectly coextensive cause. But Cartwright's General Criterion, once qualified in order to deal with the counterexamples, would automatically rule out this common-cause model for the EPR correlations.

A better response to the counterexamples is to separate the General Criterion into its sufficient and necessary parts. It then becomes clear that the qualifications are required for the sufficient part only; they are not needed

for the necessary part, which is always true. Hence, in the EPR case –in fact, in any case where a cause with coextensive operations may conceivably be at work–, it is better to enunciate these parts clearly as separate conditions, and then go on to assert the necessary condition only. In doing so, we will be in fact following Reichenbach's own example. I therefore propose that in the EPR case we suspend criterion (4.2.3), and assert criterion (4.2.2) only.

The structure that underlies the EPR correlations may well be one where the cause acts invariably to produce opposite spin outcomes in each wing, in every single instance of its operation. As we have suspended criterion (4.2.3) this common cause does not rule out a direct-link between the wings. There may well be a direct link overlaying the common cause. In fact, the quantum state itself serves perfectly well as a possible common cause of this type. For recall that, keeping the settings fixed in the same direction in both wings of the experiment, the quantum state suffices to yield the perfect correlations. The model would obviously have to be extended to take account of the fact that the full set of EPR correlations depends upon the relative orientations of the measurement apparati. This is not a problem. For, once again, if the settings are fixed, the quantum state gives the right statistics; we can therefore take the quantum state as a partial common cause of the outcomes in both wings –in each wing the corresponding apparatus setting event would be another partial cause.

Van Fraassen's 'causality' condition would, of course, be violated in this probabilistic common-cause model[25]. This shows that Van Fraassen's proof against common-cause models is inconclusive. Van Fraassen has shown that no causal theory that assumes Reichenbach's conditions is empirically adequate. But, as we have seen here, in a probabilistic causal theory the common cause obeys a much weaker condition than Reichenbach's.

[25]There is no need for either 'hidden locality' or 'hidden autonomy' to be violated too. Cartwright constructs a common cause model that obeys both these conditions in [27, appendix 1]. See also the extended discussion in [36, section 2].

4.2.4 The Empirical Adequacy of Causal Theories

The process of theory extension that takes us from screening off to the weaker criterion (4.2.2) can be put in language more akin to the semantic conception of theories. We may define any scientific theory that employs probabilistic concepts as a *causal* theory if the probability distributions defined in the theory, and the variables therein, obey a set of conditions. For instance, a causal theory may require common causes to appear in open conjunctive forks (n.b. the theory does not entail that *all* open conjunctive forks contain genuine common causes). However, several definitions of 'open conjunctive fork' are possible.

We begin by first considering Reichenbach's definition. Reichenbach defines an open conjunctive fork as a structure of three statistical variables a, b and c that obeys the conjunction of three conditions: correlation (of a and b), temporal precedence (of c with respect to a and b), and screening off (of a from b by c). We then find out that some well-established experimental correlations can not be embedded into any conjunctive fork. The EPR correlations violate the screening off condition. To restore empirical adequacy we need to extend the theory to cover the EPR correlations; and, if we want the theory to remain a *causal* theory, we need to do so by weakening the causal conditions on probability distributions. So we proceed to enlarge the set of allowed structures by weakening Reichenbach's definition; and, in particular, we withdraw the screening off condition. As a result, a larger class of open conjunctive forks is allowed. This is an extension of the causal theory because it lets more probabilistic structures in. (Notice, by the way, that we have not tinkered with the phenomena in any way.) We compare the resulting extension with the experimental correlations; as the correlations are embeddable in some of the structures allowed in this extension, the empirical adequacy of the theory is restored. If the correlations were not embeddable, we would have continued looking for further extensions of the theory, by relaxing some further causal assumptions.

The strategy that I have just described is a standard part of the normal

methodology for restoring a theory's empirical adequacy[26]. Van Fraassen[27] has described this strategy as the first of two steps in the scientific process of theory development. This first step is required to ensure empirical adequacy. The theory may be initially extended so that it embeds virtually all possible as well as actual phenomena. In order to regain informativeness, and predictive power, we then have to narrow down the set of allowed structures; this narrowing down constitutes the second step in theory-construction.

The response to Van Fraassen's arguments against causation that I have described in this chapter follows this methodology very closely. We first rejected the screening off condition, thereby admitting all probabilistic structures that satisfy Reichenbach's other two conditions. We then narrowed down the set of possible structures, by demanding that open conjunctive forks obey some other condition, such as Cartwright's General Criterion or, better still, criterion (4.2.2). Did we change our theory in this process? Is this not perhaps better described as a process of theory-change, rather than theory-development? Not at all. For all throughout we have been very careful to firmly uphold the principles of the theory –in this case, the Principle of the Common Cause, and Reichenbach's requirement that common causes occur in open conjunctive forks[28]. The principles of the theory have remained the same, but the set of permitted theoretical structures has been expanded. We have not produced a new theory, but merely *extended* the old theory.

[26]In this respect it is useful to recall the discussion of the modal interpretation in chapter 3 of this Thesis. The attempt to solve the measurement problem by relinquishing the necessary part of the *eigenstate/eigenvalue link* is formally analogous to the strategy pursued in this chapter to derive causal models of the EPR correlations. In both cases, the aim is to restore the empirical adequacy of the theory by relaxing some constraint on the available structures, thereby allowing new theoretical structures into the theory.

[27]Van Fraassen [135].

[28]In the quantum measurement case, the modal interpretation similarly sticks to the quantum theory's basic principles, including Schrödinger's equation.

In restoring the empirical adequacy of a theory, we need to expand its theoretical structures. Postulating phenomena would not help to fulfil the aim of empirical adequacy. On the contrary, when confronted with an empirically inadequate theory, postulating further phenomena can only make things worse, for it can only make the theory even more inadequate. But, as we shall see in the next section, it is however possible to solve problems in *application* by introducing further structure in the phenomena –i.e. by extending the phenomenological structures. This emphasises a fundamental methodological difference between application and confirmation.

4.3 Direct-Cause Models for EPR correlations

I have so far assumed that there is no direct causal link between the wings of an EPR experiment. Reichenbach's Principle of the Common Cause dictates that in the absence of such direct causal link there must be an underlying common cause. Cartwright's criterion (4.2.2) suggests, however, that if the common cause's operations are coextensive, an additional direct causal link between the wings cannot be ruled out, even if the joint probability for the effects is entirely accounted for by the joint expectation of the common cause's operations to produce these effects.

What then is the argument against direct causation between the wings? The standard argument is that special relativity makes a direct-causal link impossible. The acts of measurement on both particles of an EPR experiment are spacelike separated from each other. In Aspect's 1982 experiment, the setting event on the one wing is also spacelike separated from the measurement event on the other. But a causal influence between two spacelike separated events has to be transmitted at superluminal velocity –and special relativity is often taken to rule out superluminal causation.

In sections 4.3.1 and 4.3.2, I discuss the status of this standard line of argument. But I really want to concentrate on a different argument, suggested by Arthur Fine, which does not rely upon considerations of (STR). I address Fine's argument in section 4.3.3.

4.3.1 Peaceful Coexistence

Is special theory of relativity (STR) compatible with the EPR correlations? Assuming that (STR) rules out direct-cause models of the EPR correlations: do the EPR correlations nevertheless admit a direct-cause model? (STR) would be compatible with the EPR correlations only if, as a matter of principle, the EPR correlations could not admit such models. The recent debate over the 'peaceful coexistence' of (STR) and quantum theory has turned on the interpretation of the conditions of Outcome Independence (OI) and Parameter Independence (PI) described in section 4.1. Let me restate them here:

Definition 12 (Parameter Independence (PI))

$$prob(s_1/a\&b\&\lambda) = prob(s_1/a\&\lambda)$$
$$prob(s_2/a\&b\&\lambda) = prob(s_2/b\&\lambda)$$

Definition 13 (Outcome Independence (OI))

$$prob(s_1/s_2\&a\&b\&\lambda) = prob(s_1/a\&b\&\lambda)$$
$$prob(s_2/s_1\&a\&b\&\lambda) = prob(s_2/a\&b\&\lambda)$$

In section 4.1 I showed that the factorizability condition required to derive Bell's inequalities is the conjunction of (PI) and (OI). Bell's inequalities are violated by experiment; thus any empirically adequate theory must violate either (PI), or (OI), or both. I showed that quantum theory obeys (PI) and, hence, violates (OI). The burning question then, seems to be: how does (STR) fare with respect to these two conditions?

Jarrett and Shimony[29] have claimed that (STR) entails parameter independence (PI), but not outcome independence (OI). Quantum theory can then peacefully coexist with (STR): the violation of the Bell inequalities by quantum theory is due to the violation of precisely the one condition

[29] Jarrett [90], Shimony [120]. See also Ballentine and Jarrett [8].

whose failure poses no threat to (STR). Jarrett and Shimony's argument is roughly as follows. Suppose that (PI) is false, i.e. there is some state λ_1 of the particle-pair for which $P(s_1/a\&b\&\lambda_1) \neq P(s_1/a\&b'\&\lambda_1)$, for different settings b, b' of the distant apparatus (apparatus "2"). We may then be able to prepare an ensemble of particle-pairs in state λ_1. The experimenter at the "2" end can then send superluminal signals to the experimenter at "1" by carefully selecting the setting of apparatus "2". These signals are revealed at "1" as limiting frequencies in a long run of experiments on identically prepared particle-pairs in state λ_2. Different choices of the setting of apparatus "2" will result in different limiting frequencies of outcomes at "1". By appropriately selecting the settings, discernible information can be passed on between the wings. Hence we conclude that, on the assumption that (PI) is false, there exist superluminal causal influences between the wings of an EPR experiment. As special relativity is often taken to forbid such influences, it follows that a violation of (PI) entails a violation of (STR), i.e. (STR) entails (PI).

By contrast, Jarrett and Shimony argue, (STR) does not entail (OI). (OI) is false if and only if there is some possible state λ_2 of the particle-pair for which $P(s_1/s_2\&a\&b\&\lambda_2) \neq P(s_1/s_2'\&a\&b\&\lambda_2)$. However, the outcomes of a first measurement, at either wing, are stochastic. Unlike violations of (PI), the experimenter at wing "2" has no control over the relevant variable, and hence he finds it impossible to send his partner at "1" a signal[30].

Jarrett and Shimony presupose that direct causation between the wings of an EPR experiment would conflict with (STR). But, as quantum theory violates (OI) not (PI), the quantum correlations cannot be used to send signals; so in the EPR experiment, they argue, *there is no causation between the wings* –and quantum theory and special relativity can peacefully coexist.

[30]Shimony's terminology makes this very clear. He refers to the nonlocality that results from violations of (PI) as '*controllable nonlocality*', and to that which results from violations of (OI) as '*noncontrollable nonlocality*'. (Shimony [120]).

4.3.2 The Relativistic Argument Rebutted

Jarrett and Shimony's argument, however, –as pointed out by a number of authors–, is flawed. I want to take issue with their argument in three different stages; at each one I raise an objection, of increasing strength. First, I want to take issue with the claim that superluminal causation requires superluminal signalling. Second, I want to lessen the import of parameter independence and outcome independence by showing that, as conditions on putative causal factors, (PI) and (OI) are arbitrary: other choices are also available. Finally, I take issue with the claim that (STR) entails (PI). As a matter of fact (STR) does not strictly entail (PI). More precisely: (STR) does not entail (PI) anymore than it entails (OI) –in either case, an additional assumption is required.

I begin by pointing out that nothing in the concept of probabilistic causation requires a causal connection to always be a vehicle for signalling. Recall that in probabilistic theories of causation a cause C at some space-time point c may act genuinely stochastically to produce an effect A at some other point a. Suppose that C raises the probability of A, although it does not make it certain: we will then say that C is a (probabilistic) cause of A. Suppose also that we control for C and that we have a large number of C-tokens available. The causal relation can then be employed to send a signal from c to a, by raising the frequencies of A's at a. But C itself may be a stochastic effect of some further cause D, making it impossible for us to control for C. No matter: we are surely not going to withdraw our assertion that C is a cause of A because of that! For C's capacity to bring about A remains unaltered, as it is indeed exhibited whenever C occurs –and that surely is the content of the causal claim that we have been making all along. Thus, causation cannot generally be analysed in terms of signalling. The fact that some causal connections cannot be employed for signalling does not make them any less causal. In general, causation does not require the kind of controllable signalling demanded by Jarrett and Shimony. There may well be causation between the wings, even if (OI), rather than (PI), fails.

The violation of (OI) is then no less problematic than the violation of (PI). We are in trouble either way, if it is true that (STR) forbids superluminal causal influences. This, however, is not a very telling objection to Jarrett and Shimony's argument. For in fact, special relativity theory does not forbid superluminal causation at all. The point has been convincingly argued by Tim Maudlin[31], and I shall not rehearse it in detail here. To cut a long story short: the Lorentz group of (STR) lays out no constraints whatever on the speed of causal connections. Instead (STR) dictates a constraint on the acceleration of material moving objects: no infraluminally moving object can be accelerated past the speed of light; and viceversa, no superluminally moving object can be deaccelerated to within light's speed. As causal connections, however, need not be associated with the trajectories of moving material objects, nor do they need to be 'carried' by such objects, (STR) cannot be said to lay out constraints upon the speed of causal connections.

So where does the argument that (STR) rules out superluminal causation originate from? As it happens, the argument originates in the notorious 'signalling' paradoxes of (STR): in certain spacetime contexts, it becomes possible to send signals to one's own past; such signals may (although they need not) give rise to logical contradictions. (I may be able to signal to my own past, and instruct myself not to send the signal in the first place: so I don't send a signal if and only if I do.) Thus, *what we meant all along* is that (STR) forbids certain kinds of superluminal *signals*. But I have already established that causation does not require signalling. (STR) may then be said to admit superluminal causation of the kind prescribed by the failure of (OI), while ruling out some subclass of the superluminal signals allowed by violations of (PI). As quantum theory violates (OI) but not (PI), the quantum correlations cannot be used for any form of superluminal signalling between the wings; and hence, *a fortiori* they cannot be used to send any of the special kind of signals that (STR) would find objectionable. Jarrett and Shimony's argument for peaceful coexistence is now again on its feet[32].

[31] Maudlin [99].

[32] But notice that, in this revised form, the argument can shed no light on the *causal*

The second objection to Jarrett and Shimony's argument is that (PI) and (OI) are not the only conditions whose conjunction is logically equivalent to factorizability. Tim Maudlin[33] has noted that the following two alternative conditions, which we may call (PI') and (OI'), also yield factorizability:

Definition 14 (PI')

$$prob\,(s_1/s_2\&a\&b\&\lambda) = prob\,(s_1/s_2\&a\&\lambda)$$
$$prob\,(s_2/s_1\&a\&b\&\lambda) = prob\,(s_2/s_1\&b\&\lambda)$$

Definition 15 (OI')

$$prob\,(s_1/s_2\&a\&\lambda) = prob\,(s_1/a\&\lambda)$$
$$prob\,(s_2/s_1\&b\&\lambda) = prob\,(s_2/b\&\lambda)$$

The first condition states that the probability of a given result in one wing of the experiment, given that one already knows the result on the other, does not depend upon the actual setting of the distant apparatus. We may be told what the result of the distant measurement was –'up', or 'down'–, while being kept in the dark as to the setting of the distant apparatus when this measurement was carried out. Condition (PI') says that the probabilities for outcomes on our wing are not affected when we receive extra information concerning the distant setting. So (PI') is the analogue of parameter independence: both make the information about the distant setting irrelevant to the assessment of the probabilities of measurement outcomes.

The second condition, (OI'), states that the probabilities for outcomes at one wing do not depend on the outcomes of measurements performed in the distant wing. So (OI') is analogous to outcome independence. But notice

processes that may be operating in an EPR experiment, nor can it serve to rule out causal models for the EPR correlations. The argument now applies to signals, not to causal connections.

[33]Maudlin [99, pp 93-99].

that, unlike its original counterpart, (OI') does not require conditionalising on, and keeping fixed, the distant setting.

In construing their original conditions, (PI) and (OI), as conditions on causal influences, Jarrett and Shimony took it that if some factor had no statistical significance, then it had no causal significance. Thus (PI) could be construed as asserting that the settings on one wing cannot causally influence the outcomes on the other wing, while (OI) would assert that outcomes on one wing cannot bring about outcomes on the other. The same interpretation can of course be given to Maudlin's analogous (PI') and (OI') conditions. (PI') would assert that distant settings do not causally influence the outcomes, while (OI') would assert that distant outcomes have no causal influence on nearby outcomes. But now, the situation is strikingly reversed: quantum theory violates (PI'), while satisfying (OI'). (It is straightforward to show that quantum theory satisfies (OI'): in the absence of any information about the distant setting, the information regarding the actual outcome, 'up' or 'down', of the distant measurement can have no effect on the probabilities for outcomes on the nearby wing.)

So, by applying to Maudlin's (PI') and (OI') conditions the very same causal interpretation often applied to (PI) and (OI), it turns out that quantum theory exhibits causal links between settings and outcomes and no causal influences between outcomes. Jarrett and Shimony's verdict is dramatically changed, and peaceful coexistence turns into open war. For, on the very reading of (STR)'s constraints that favours (OI) over (PI), the violation of (PI') takes quantum theory into straightforward conflict with (STR). Quantum theory now violates precisely the one condition whose failure allows superluminal signalling of the sort we have taken (STR) to prohibit.

The final, and most decisive, objection to Jarrett and Shimony's argument is due to Jeremy Butterfield[34]. Butterfield shows that (STR) does not strictly entail (PI). A further assumption is required. Unfortunately for Jarrett and Shimony's argument, given this assumption (STR) entails (OI)

[34] Butterfield [22]. See also Jones and Clifton [92] for a related criticism.

as well. Hence, as regards the special theory of relativity (OI) and (PI) are on a par: together with the further assumption, (STR) entails both; without the further assumption, it entails neither.

Butterfield begins by noting that both (OI) and (PI) are essentially screening off conditions. On the one hand (OI) asserts that the conjunction of a, b and λ screens off s_1 from s_2. The settings and the state together screen off the outcomes from each other. On the other hand, (PI) contains two screening off conditions: first, the conjunction of a and λ screens off s_1 from b; second, b and λ screen off s_2 from a. That is, the setting on the one wing together with the state, screen off the outcome on that wing from the setting in the other.

In section 4.2 I argued that screening off is not a necessary condition on common causes. Hence, for instance, taking the conjunction of a, b and λ to be the common cause of s_1 and s_2 won't justify (OI) –if the causal relation is genuinely probabilistic, the conjunction of a, b and λ won't normally screen off. Similarly for (PI): taking $a\&\lambda$ ($b\&\lambda$) as the common cause of s_1 and b (s_2 and a) gives no grounds to uphold (PI). Butterfield suggests that we take the different conjunctions of a, b and λ to constitute indexes for spacetime regions, and that we take those regions to constitute the common past of the events that get screened off. He then enunciates the following extension of Reichenbach's Principle of the Common Cause[35]:

Definition 16 (Past Prescribes Stochastic Independence (PPSI)) *If two events are correlated, but one does not cause the other, then the probability distribution prescribed by the total physical state of their common past makes them stochastically independent (and so [...] screens them off).*

This principle restores screening off as a necessary condition, not on common causes, but on the total state of the common past of two correlated, and not directly causally connected, events. For instance, as regards the correlated outcomes s_1, s_2 on the wings of an EPR experiment, (PPSI) implies that

[35]Butterfield [22, page 123].

either there is a direct causal link between these events or the common past of s_1 and s_2 screens them off from each other, as prescribed by (OI). If we now take (STR) to rule out direct causal influences between the wings, we can conclude that the common past screens off, and hence (OI) holds: (STR) and (PPSI) together entail (OI).

There are two obvious rubs. First: why should the common past screen off? This is a subtle issue, and I shall not attempt to treat it in detail here. Butterfield makes a compelling case for (PPSI), by showing that in the standard macroscopic and classical examples of failure of screening off by the putative common cause, there is a region of spacetime that satisfies (PPSI). However, quantum theory violates (OI) and, if superluminal causation is impossible, (PPSI) must be false in the EPR experiment. So, Butterfield claims, the fact that (PPSI) fails in the EPR experiment is a truly surprising one, and it shows just how deep the revolutionary implications of quantum theory go. But here comes the second rub: didn't we agree that (STR) does not rule out superluminal causation? Fortunately it doesn't matter much what interpretation of the relativistic constraints we take here. My purpose, following Butterfield, is to show that (PI) is no more entailed by (STR) –on its own– than is (OI). We may grant, for the sake of argument, the stronger interpretation of (STR) as ruling out superluminal causation. It won't affect our conclusions. For if (STR) fails to entail (PI) on this interpretation, it can hardly entail it on *weaker* interpretations.

And indeed (STR) fails to entail (PI). But not if (PPSI) is also assumed. Recall what (PI) says: $P(s_1/a\&b\&\lambda) = P(s_1/a\&\lambda)$; the measurement outcome at '1' exhibits no statistical dependence upon the distant setting at '2' (and similarly for the outcome at '2' and the setting at '1', but let me consider one case only). Recall also that in Aspect's experiment the outcomes are spacelike separated from the distant settings. If we then take (STR) to rule out superluminal causation, (PI) turns into yet another screening off condition. As we noted earlier, there is no reason to expect (PI) to hold, even if we take a and λ to be the common cause of s_1 and b. There is no reason to expect the correlations to wash out conditional on the common

cause, even in the absence of any direct-cause link. So (STR) does not entail (PI). To enforce screening off we need to appeal to a principle like (PPSI): only then can we expect the correlations between s_1 and b to vanish when we conditionalise on a and λ. Hence, (PI) and (OI) are on equal grounds with respect to (STR). Neither is entailed by (STR) on its own; and both are entailed by the conjunction of (STR) and (PPSI). And this is good news for the peaceful coexistence of quantum theory and special theory of relativity: the failure of (PI) spells no more disaster for peaceful coexistence than does the failure of (OI)[36].

One final remark. It is now possible to see why I have taken Butterfield's to be a stronger objection than Maudlin's. Although Maudlin's objection makes Jarrett and Shimony's argument less plausible, I do not think it is a decisive objection. For Maudlin's conditions are not really on a par with parameter and outcome independence. The latter two can be justified by means of Butterfield's (PPSI): in the absence of superluminal causation between the wings, and given (PPSI), both (OI) and (PI) hold. But, (PPSI) cannot be similarly used to make (PI') plausible. Recall that (PI') says that $P(s_1/b\&s_2\&a\&\lambda) = P(s_1/s_2\&a\&\lambda)$, i.e., read as a screening off condition, the conjunction of s_2, a and λ screens off s_1 from b. But it is not possible to construe s_2, a, λ as indexes for the common past of s_1 and b: for *in all inertial frames of reference* b is in the past of s_2! No choice of spacetime region is then available to justify Maudlin's (PI') condition by means of (PPSI).

4.3.3 A Quantum Mechanical Model

Suppose that there is superluminal causation between the wings of the EPR experiment. What kind of information needs to be transmitted? A direct-

[36]Rather than rejecting (STR), it is always perfectly possible to put the blame, as Butterfield does, on (PPSI)'s doorstep. On the other hand, *we are not forced* to reject (PPSI) either. We may instead choose to put the blame on the strong interpretation of (STR). This is a plausible way out too, –for it is indeed dubious that (STR) prohibits superluminal causation. I take this line in section 4.3.3.

cause model would have to transmit two types of information directly from one wing to the other: information regarding the outcome of a particular measurement on a given wing, and information regarding the setting of the measurement apparatus on that wing. All the information could be transmitted in one go, at the conclusion of the first measurement interaction on either particle. But a model in which all the information is transmitted in one go, says the critic, would be against the spirit of both relativity and quantum mechanics. For in that case, information regarding the setting of the measurement apparatus on the one wing is essential for an accurate calculation of the probabilities for outcomes on the distant wing. In this final section, I shall be primarily concerned with the claim that a direct-cause model would be in conflict with the spirit of quantum theory. This claim has been made explicitly by Arthur Fine[37]. What would be required, asks Fine, to show that there must exist a direct causal link between the wings of an EPR experiment? Well, if the denial of the existence of direct links *entailed* Bell's inequalities, it would follow from the experimental violation of the inequalities that there must be such links. Fine then looks for a 'locality' principle, the violation of which would entail the existence of superluminal influences between the wings. He does not give us a specific principle, but he describes part of what such a principle would have to state. Whatever its precise content turns out to be, this principle would explicitly deny that there are any sort of causal influences between the wings of an EPR experiment; it would deny, for instance that the outcomes of measurements on a given wing causally depend upon spacelike separated events in the other wing. Thus, if we could show that the Bell inequalities are entailed by this principle we would have achieved our aim: we would have shown that there must exist causal influences between the wings. On the other hand, as Fine notes, if it can be shown that the principle is consistent with the *denial* of the Bell inequalities, our argument would collapse –and we would then never be in a position to show that there *must* exist superluminal causal influences

[37] Fine [65, pp 183-185].

between the wings.

According to Fine the quantum theory itself shows that the 'locality' principle is consistent with the denial of Bell's inequalities. He writes:

> For [the locality principle] *is built into* the quantum theory, according to which there is no influence between the two wings of the experiment, that is, no physical interaction of the sort that is represented by terms in the Hamiltonian of the composite system at the time one or the other component is measured. (Fine [65, page 183], my italics).

Fine's claim is that quantum theory fully subscribes to, and in fact embodies, the 'locality' principle. And yet, the theory predicts a violation of Bell's inequalities. So the quantum theory itself provides a context in which our 'locality' principle is consistent with the denial of Bell's inequalities. But then our 'locality' principle cannot *entail* the Bell inequalities, and we are left with no positive argument for a direct-cause link between the wings. (This does not mean that we have *ruled out* direct-cause links, but rather that we have no hope ever of providing a positive argument for them.)

Note first of all, that Fine's argument works entirely within quantum theory, and makes no appeal to relativity. Fine claims that the quantum theory contains no resources to model direct causation between the wings. Indeed, for Fine's argument to work, there must exist no application of the quantum theory to the putative phenomenon of direct-causation between the wings. Fine thinks that the absence of interaction terms in the Hamiltonian of the composite system gives conclusive reasons to believe that there is no quantum treatment of direct-cause links.

Can there be no quantum treatment of direct-cause links? I shall now outline the very basic elements of a direct-cause model for the EPR correlations built together with Nancy Cartwright[38]. The model postulates *carriers* which transmit causal influence between the wings. Each particle has an associated carrier, with a quantum state of its own. When a measurement is

[38] Cartwright and Suárez [35].

performed on one particle, a stochastic outcome results, with corresponding Born probability; and the associated carrier is released. This carrier attaches itself to the partner particle in the opposite wing. The dynamics for the interaction between a particle and a carrier is governed by a specific rule of evolution of the particle/carrier composite system, which we have to add to the quantum formalism. This dynamics is unitary, so the evolution of the particle/carrier system is purely quantum mechanical –subsequent measurements on the partner particle will not result in stochastic outcomes, but will reveal spin values possessed by the particle before measurement. This model replicates the quantum mechanical statistics, including the conditional probabilities at the heart of the EPR correlations.

Is this model ruled out by special relativity? I do not think so. In this model the transmission of causal influence is instantaneous between the wings in the rest frame of the laboratory. This gives us two options. The first is outright conflict with (STR). This would be the case if the process of causal transmission was used to pick out a preferred frame. We could then assert that the process of causal transmission in an EPR experiment is always instantaneous in some particular, privileged frame, and decline (STR)'s invitation to consider the physical situation from the point of view of any other inertial frames. The second option is to insist that relativity must hold. The principle of relativity then requires that no differences in the physical laws or mechanisms in this situation must result from mere changes of inertial frame description. All physical laws and mechanisms should be the same from the point of view of all Lorentz invariant frames. Suppose that we make a measurement on particle "2" first. Consider the frame in which particle "1" is at rest (this is a valid inertial frame only if the first particle travels at infraluminal speed, of course, i.e. if it is not a photon.) As described in that frame, the causal transmission is not instantaneous, but has finite superluminal speed. There is no conflict with the principle of relativity, though: the order of events is unchanged, and so are the mechanisms that generate the correct EPR statistics. However, consider now the same measurement process as described in the rest frame of particle "2".

The carrier is received at particle "1"s end *before* it is ejected as a result of a measurement on particle "2": the effect precedes the cause[39]. The direct-cause model is committed to the cogency of backwards causation. This is perhaps not surprising. Indeed this is also a consequence of other causal models of the EPR correlations –such as, for instance, the zigzag model of Costa de Beauregard[40]. No matter: conflict with (STR) is nonetheless averted.

The direct-cause model constitutes a further illustration of the profound differences between the structure of application and the structure of empirical adequacy. To resolve difficulties with empirical adequacy, we expand the set of structures that the theory makes available; adding structure to the phenomena seems hopeless in this case. But to resolve conflicts of application, adding phenomena may well help, as is the case with this direct-cause model of EPR. Instead of relaxing some of the constraints on the quantum theory to let new structures in –such as for instance introducing new interaction terms in the Hamiltonian of the composite–, the model postulates further phenomenological structure, in the form of *carriers* that transmit the causal influence between the wings. These carriers are modelled as quan-

[39]A reinterpretation principle is also generally available for tachyons, but I do not think that it works here. The reinterpretation principle states that a tachyon travelling backwards, i.e. one emitted in the future and received in the past, can always be interpreted as a tachyon with negative energy travelling forwards in time. The problems usually associated with the reinterpretation principle have to do with negative energies: any system could increase its energy infinitely by emitting infinitely many negative energy tachyons. However, the direct-cause model does not assume that carriers must transport energy, so this problem does not arise. But there is another problem. One would have to suppose that it was particle "1" which originally emitted the carrier, rather than particle "2". But one of the assumptions of this thought-experiment is that the measurement is made on particle "2" not "1". Hence, according to the description in the rest frame of "2" the carrier is ejected by particle "1" even if no measurement is ever made on it. This would constitute a violation of the relativity principle: some physical process (spontaneous carrier-release) is describable in rest-frame "2" but not in "1".

[40]Costa de Beauregard [42].

tum objects themselves, and they are ascribed a quantum state. It is then possible to provide a theoretical law that represents the physical interaction between particles and carriers.

The overall methodology in this case of theory-application is as follows. First we postulate further phenomenological structure –hence making the theory even more grossly empirically inadequate. But it then becomes possible to 'adapt' the theory, by introducing a new operational law. From the point of view of the theory, the introduction of this new law looks *ad hoc* –it looks as a device for simply saving the phenomena–; and hence this move yields no confirmatory boost for the theory. But from the point of view of the newly postulated phenomena, the introduction of this law is justified by the fact that the carriers are modelled as quantum objects. It then becomes possible to apply the quantum theory to the EPR experiment, in order to provide a direct-cause model of the correlations.

4.4 Summary

In this chapter, I have focused on some more typical quantum phenomena, namely the correlations between spin values of entangled pairs of particles first described by Einstein, Podolsky and Rosen. In the first section, in order to set the scene, I review the EPR correlations, and I introduce the conditions of factorizability, outcome and parameter independence. In the second section, I address Reichenbach's Principle of the Common Cause and Van Fraassen's allied argument against common-cause models of the EPR correlations. I argue, following Cartwright, that Van Fraassen's analysis does not cut ice against genuinely probabilistic common-cause models of the EPR correlations. For, in those cases, Reichenbach's conditions are too limited and need to be appropriately extended. And although the EPR correlations fail to fit the original Reichenbach conditions, they fit the extended conditions for probabilistic causes very accurately; a common-cause model of the EPR correlations is then possible. In the third and final section, I look at the arguments against direct-cause models of the EPR correlations. I consider

briefly, and reject, arguments from relativity theory. Then I take on Arthur Fine's objection to direct-cause models. Fine has argued that the quantum theory itself rules out such direct-cause explanations: no consistent application of the quantum theory can be found that describes such direct-causal links between the EPR particles. I argue that, to the contrary, it is possible to construct a quantum mechanical model of the putative direct-cause links operating in the EPR experiment.

Bibliography

[1] D. ALBERT, *Quantum Mechanics and Experience*, Harvard University Press, 1993.

[2] D. ALBERT AND B. LOEWER, *Wanted dead or alive: Two attempts to solve schrodinger's paradox*, PSA, 1 (1990), pp. 277–285.

[3] ——, *The measurement problem: Some "solutions"*, Synthese, 86 (1991), pp. 87–98.

[4] ——, *Non-ideal measurements*, Foundations of Physics Letters, 6 (1993), pp. 297–303.

[5] A. ASPECT, J. DALIBARD, AND G. ROGER, *Experimental realization of Einstein-Podolski-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities*, Phys. Rev. Letts., 49 (1982), pp. 1804–1807.

[6] G. BACCIAGALUPPI AND M. DICKSON, *Modal interpretations with dynamics.* In Preparation, 1996.

[7] G. BACCIAGALUPPI AND M. HEMMO, *Modal interpretations of imperfect measurements.* Studies in History and Philosophy of Physics, 1996.

[8] L. E. BALLENTINE AND J. JARRETT, *Bell's theorem: Does quantum mechanics contradict relativity?*, American Journal of Physics, 55 (1987).

[9] J. BARDEEN, *Theory of superconductivity*, (1959), pp. 274–365.

[10] J. BARDEEN, L. COOPER, AND J. SCHRIEFFER, *Theory of Superconductivity*, Phys. Rev., 108 (1957), p. 1175.

[11] R. BECKER, F. SAUTER, AND G. HELLER, Z.Physik, 85 (1933), p. 772.

[12] J. BELL, *On the Einstein Podolsky Rosen paradox*, Physics, 1 (1964), pp. 195–200. Reprinted in [139].

[13] B. BLEANEY AND B. BLEANEY, *Electricity and Magnetism*, Oxford University Press, Oxford, third ed., 1976.

[14] J. BOGEN AND J. WOODWARD, *Saving the phenomena*, The Philosophical Review, 97 (1988), pp. 303–352.

[15] D. BOHM, *Quantum Theory*, Prentice Hall, Englewood Cliffs, NJ, 1951.

[16] R. BOYD, *Realism, underdetermination and a causal theory of evidence*, Nous, 7 (173), pp. 1–12.

[17] ——, *The current status of scientific realism*, in Leplin [96], pp. 41–82.

[18] H. BROWN, *The insolubility proof of the quantum measurement problem*, Foundations of Physics, 16 (1986).

[19] H. BROWN, M. SUÁREZ, AND G. BACCIAGALUPPI, *Are 'sharp values' of observables always objective elements of reality?*, in The Modal Interpretation of Quantum Mechanics, D. Dieks and P. Vermaas, eds., Kluwer, Forthcoming in 1998.

[20] O. BUENO, *Empirical adequacy: A partial structures approach.* Leeds University preprint, 1995.

[21] P. BUSCH, P. LAHTI, AND P. MITTELSTAEDT, *The Quantum Theory of Measurement*, Springer-Verlag, Berlin, 1991.

[22] J. BUTTERFIELD, *A space-time approach to the Bell inequality*, in Cushing and McMullin [39], pp. 114–144.

[23] R. CARNAP, *The Logical Syntax of Language*, Harcourt, New York, 1937.

[24] ——, *Philosophy and Logical Syntax*, Thoemmes Press, Bristol, 1996. Originally published in 1935.

[25] N. CARTWRIGHT, *A dilemma for the traditional interpretation of quantum mixtures*, in Proceedings of the Philosophy of Science Association 1972 Meeting, K. Schaffner and R. Cohen, eds., Dordrecht, 1974, Reidel, pp. 251–258.

[26] ——, *How the Laws of Physics Lie*, Oxford University Press, Oxford, 1983.

[27] ——, *Nature's Capacities and Their Measurement*, Oxford University Press, Oxford, 1987.

[28] ——, *How to tell a common cause: Generalizations of the conjunctive fork criterion*, in Probability and Causality, J. Fetzer, ed., Reidel, 1988, pp. 181–188.

[29] ——, *The Born-Einstein debate: Where application and explanation separate*, Synthese, 81 (1989), pp. 271–282.

[30] ——, *Capacities and abstractions*, in Scientific Explanation, P. Kitcher and W. Salmon, eds., Minnesotta Studies in the Philosophy of Science, University of Minnesotta Press, Minneapolis, 1989.

[31] ——, *How we relate theory to observation*, in Horwich [88], pp. 259–274.

[32] ——, *Marks and probabilities: Two ways to find causal structure*, in Yearbook 1/93, Institute Vienna Circle, F. Stadler, ed., Dordrecht-Boston-London: Kluwer, 1993.

[33] ——, *False idealization: A philosophical threat to scientific method*, Philosophical Studies, 77 (1995), pp. 339–352.

[34] N. CARTWRIGHT, T. SHOMAR, AND M. SUÁREZ, *The tool-box of science*, in Proceedings of the IUHPS Warsaw Conference, Poznan Studies in the Philosophy of Science and Humanities, Rodopi, 1994.

[35] N. CARTWRIGHT AND M. SUÁREZ, *A causal model for EPR.* Preprint.

[36] H. CHANG AND N. CARTWRIGHT, *Causality and realism in the EPR experiment*, Erkenntnis, 38 (1993), pp. 169–189.

[37] P. CHURCHLAND AND C. HOOKER, eds., *Images of Science: Essays on Realism and Empiricism*, Chicago University Press, Chicago, 1985.

[38] N. D. COSTA AND S. FRENCH, *The model-theoretic approach in the philosophy of science*, Philosophy of Science, 57 (1990), pp. 248–265.

[39] J. CUSHING AND E. MCMULLIN, eds., *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem*, University of Notre Dame Press, 1989.

[40] P. DAHL, *Superconductivity: Its Historical Roots and Development from Mercury to the Ceramic Oxides*, American Institute of Physics, 1992.

[41] M. DALLA CHIARA AND G. TORALDO DI FRANCIA, *Formal analysis of physical theories*, in Problems in the Foundations of Physics, Proceedings of the International School of Physics 'Enrico Fermi', 1979.

[42] C. DE BEAUREGARD, *Time symmetry and the Einstein paradox*, Nuovo Cimento B, 42 (1977), pp. 41–64.

[43] M. DELSETA AND M. SUÁREZ, *The physical impossibility of ideal measurements.* Forthcoming in the Proceedings of the Logic, Methodology and Philosophy of Science Division of the IUHPS conference, August 1995.

[44] B. D'ESPAGNAT, *Conceptual Foundations of Quantum Mechanics*, Benjamin, Reading, Massachusetts, 1976.

[45] D. DIEKS, *Resolution of the measurement problem through decoherence of the quantum state*, Physics Letters A, 142 (1989), pp. 439–446.

[46] ——, *On some alleged difficulties in the interpretation of quantum mechanics*, Synthese, 86 (1991), pp. 77–86.

[47] ——, *The modal interpretation of quantum mechanics, measurements and macroscopic behavior*, Physical Review D, 49 (1993), pp. 367–371.

[48] P. DOWE, *Wesley Salmon's process theory of causality and the conserved quantity theory*, Philosophy of Science, 59 (1992), pp. 195–216.

[49] P. DUHEM, *The Aim and Structure of Physical Theory*, Princeton University Press, Princeton, New Yersey, 1954.

[50] ——, *To Save the Phenomena*, The University of Chicago Press, 1969.

[51] J. EARMAN AND A. SHIMONY, *A note on measurement*, Nuovo Cimento, 54B (1968).

[52] A. EINSTEIN, B. PODOLSKY, AND N. ROSEN, *Can quantum-mechanical description of the physical reality be considered complete*, Phys. Rev., 47 (1935), pp. 777–780.

[53] H. EVERETT, *"Relative state" formulation of quantum mechanics*, Reviews of Modern Physics, 29 (1957), pp. 454–62. Reprinted in [139].

[54] U. FANO, *Description of states in quantum mechanics by density matrix and operator techniques*, Reviews of Modern Physics, (1957), pp. 74–93.

[55] P. K. FEYERABEND, *On the quantum theory of measurement*, in Observation and Interpretation, S. Korner, ed., Academic Press, New York, 1957, pp. 121–130.

[56] A. FINE, *On the general quantum theory of measurement*, Proc Camb Phil Soc, 65 (1969), pp. 111–122.

[57] ——, *Insolubility of the quantum measurement problem*, Physical Review D, 2 (1970).

[58] ——, *Probability and quantum mechanics*, British Journal for the Philosophy of Science, 24 (1973).

[59] ——, *The two problems of quantum measurement*, in Logic, Methodology and Philosophy of Science, P. Suppes, ed., vol. IV, North-Holland, 1973, pp. 567–581.

[60] ——, *Hidden variables, joint probability, and the Bell inequalities*, Physical Review Letters, 48 (1982), pp. 291–295.

[61] ——, *Some local models for correlation experiments*, Synthese, 50 (1982), pp. 279–294.

[62] ——, *The Shaky Game: Einstein, Realism and the Quantum Theory*, Chicago University Press, Chicago, 1986.

[63] ——, *Unnatural attitudes: Realist and instrumentalist attachments to science*, Mind, 95 (1986), pp. 149–179.

[64] ——, *With complacency or concern: Solving the quantum measurement problem*, in Kelvin's Baltimore Lectures and Modern Theoretical Physics: Historical and Philosophical Perspectives, MIT Press, Cambridge, Massachusetts, 1987, pp. 491–505.

[65] ——, *Do correlations need to be explained?*, in Cushing and McMullin [39], pp. 175–194.

[66] ——, *Measurement and quantum silence*, in Correspondence, Invariance and Heuristics: Essays for Heinz Post, Kluwer, Dordrecht, 1992?, pp. 279–294.

[67] ——, *Resolving the measurement problem: Reply to Stairs*, Foundations of Physics Letters, 5 (1992).

[68] S. FRENCH AND J. LADYMAN, *Superconductivity and structures: Revisiting the london account*, Studies in the History and Philosophy of Modern Physics, (Forthcoming).

[69] M. FRIEDMAN, *Review of The Scientific Image*, The Journal of Philosophy, (1982), pp. 274– 283.

[70] ——, *Foundations of Space-Time Theories*, Princeton University Press, Princeton, 1983.

[71] K. GAVROGLU, *Fritz London: A Scientific Biography*, Cambridge University Press, Cambridge, 1995.

[72] G. C. GHIRARDI, A. RIMINI, AND T. WEBER, *Unified dynamics for microscopic and macroscopic systems*, Physical Review D, 34 (1986).

[73] R. GIERE, *Explaining Science: A Cognitive Approach*, University of Chigago Press, 1988.

[74] ——, *The cognitive structure of scientific theories*, Philosophy of Science, 61 (1994).

[75] N. GISIN, *Quantum measurements and stochastic processes*, Physics Review Letters, 52 (1984), pp. 1657–1660.

[76] H. GOLDSTEIN, *Classical Mechanics*, Addison-Wesley, second ed., 1980.

[77] GORTER, *Theory of superconductivity*, Nature, 132 (1933), p. 931.

[78] GORTER AND CASIMIR, *On supraconductivity I*, Physica, 1 (1934), pp. 306–320.

[79] I. HACKING, *Language, truth and reason*, in Hollis and Lukes [87], pp. 48–66.

[80] ——, *Representing and Intervening*, Cambridge University Press, 1983.

[81] H. E. HALL, *Solid State Physics*, John Wiley and Sons, 1974.

[82] R. HEALEY, *The Philosophy of Quantum Mechanics: An Interactive Interpretation*, Cambridge University Press, Cambridge, 1989.

[83] C. G. HEMPEL, *Aspects of Scientific Explanation*, Free Press, New York, 1965.

[84] ——, *Philosophy of Natural Science*, Prentice Hall Inc, Englewood Cliffs, New Jersey, 1966.

[85] ——, *On the 'standard' conception of scientific theories*, in Minnesota Studies in the Philosophy of Science, M. Radner and S. Winokur, eds., vol. VI, University of Minnesota Press, Minnesota, 1972, pp. 142–163.

[86] ——, *Formulation and formalization of scientific theories*, in Suppe [124], pp. 244–265.

[87] M. HOLLIS AND S. LUKES, eds., *Rationality and Relativism*, MIT Press, Cambridge, Massachusetts, 1982.

[88] P. HORWICH, ed., *World Changes: Thomas Kuhn and the Nature of Science*, MIT Press, Cambridge, Massachusetts, 1993.

[89] R. I. G. HUGHES, *The Structure and Interpretation of Quantum Mechanics*, Harvard University Press, 1989.

[90] J. JARRETT, *On the physical significance of the locality conditions in the Bell arguments*, Nous, 18 (1984), pp. 569–589.

[91] ——, *Does Bell's theorem apply to theories that admit time-dependent states?*, in New Techniques and Ideas in Quantum Measurement Theory, D. Greenberger, ed., Annals of the New York Academy of the Sciences, 1986, pp. 428–437.

[92] M. JONES AND R. CLIFTON, *Against experimental metaphysics*, in The Philosophy of Science, vol. 18 of Midwest Studies in Philosophy, University of Notre Dame Press, 1993.

[93] M. KAISER, *From rocks to graphs: The shaping of phenomena*, Synthese, 89 (1991), pp. 111–133.

[94] C. KITTEL, *Introduction to Solid State Physics*, John Wiley and Sons, 1953.

[95] S. KOCHEN, *A new interpretation of quantum mechanics*, in Symposium in the Foundations of Modern Physics, P. Lahti and P. Mittelstaedt, eds., World Scientific, Singapore, 1985.

[96] J. LEPLIN, ed., *Scientific Realism*, University of California Press, Berkeley, 1984.

[97] F. LONDON, *Macroscopical interpretation of supraconductivity*, Proceedings of the Royal Society, A152 (1935), pp. 24–34.

[98] F. LONDON AND H. LONDON, *The electromagnetic equations of the supraconductor*, Proceedings of the Royal Society, A149 (1934), pp. 71–88.

[99] T. MAUDLIN, *Quantum Non-locality and Relativity*, Blackwell, 1994.

[100] E. MCMULLIN, *Galilean idealization*, Studies in History and Philosophy of Science, 16 (1985), p. 264.

[101] W. MEISSNER AND R. OCHSENFELD, Naturwissenschaften, 11 (1933), p. 787.

[102] M. MORGAN AND M. MORRISON, eds., *Models in Physics and Economics*, Cambridge University Press, Cambridge, forthcoming, 1998.

[103] M. MORRISON, *Unification, realism and inference*, British Journal for the Philosophy of Science, 41 (1990), pp. 305–332.

[104] ——, *Mediating models: between physics and the physical world*, Philosophia Naturalis, (forthcoming).

[105] A. MORTON, *Mathematical models: Questions of trustworthiness*, British Journal for the Philosophy of Science, 44 (1993).

[106] F. MÜLLER, M. SUÁREZ, AND P. VERMAAS, *Some remarks on the relation between the time-evolution of composite systems and their subsystems in quantum mechanics.* To be submitted to Physics Letters A, 1997.

[107] E. NAGEL, *The Structure of Science: Problems in the Logic of Scientific Explanation*, Harcourt, Brace and World, Inc, New York, 1961.

[108] N. NERSESSIAN, ed., *The Process of Science*, Martinus Nijhoff, 1987.

[109] J. V. NEUMANN, *Mathematical Foundations of Quantum Mechanics*, Princeton, 1932.

[110] K. ONNES, *Commun. kamerlingh onnes lab.*, Tech. Rep. 34b, Univ. Leiden, 1913.

[111] H. POST, *Correspondence, invariance and heuristics*, Studies in History and Philosophy of Science, 2 (1971), pp. 213–255.

[112] M. PRZELECKI, *The Logic of Empirical Theories*, Routledge, London, 1969.

[113] H. PUTNAM, *Mathematics, Matter and Method*, Cambridge University Press, Cambridge, 1975.

[114] M. REDHEAD, *Models in physics*, British Journal for the Philosophy of Science, 31 (1980).

[115] ——, *Incompleteness, Non-locality and Realism*, Oxford University Press, Oxford, 1987.

[116] H. REICHENBACH, *The principle of anomaly in quantum mechanics*, Dialectica, 2 (1948), pp. 337–50.

[117] H. REICHENBACH, *The Direction of Time*, University of California Press, 1956.

[118] W. SALMON, *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, 1984.

[119] ——, *Causality without counterfactuals*, Philosophy of Science, 61 (1994), pp. 297–312.

[120] A. SHIMONY, *Controllable and uncontrollable nonlocality*, in Proceedings of the International Symposium: Foundations of Quantum Mechanics in the Light of New Technology, S. Kamefuchi, ed., 1984, pp. 225–230.

[121] A. SHIMONY, *Events and processes in the quantum world*, in Quantum Processes in Space and Time, Oxford University Press, 1986, pp. 182–203.

[122] L. SKLAR, *Topology vs measure in statistical mechanics*. Manuscript, 1997.

[123] M. SUÁREZ, *On the physical impossibility of ideal quantum measurements*, Foundations of Physics Letters, 9 (1996).

[124] F. SUPPE, ed., *The Structure of Scientific Theories*, University of Illinois Press, Urbana and Chicago, 1977.

[125] P. SUPPES, *A comparison of the meaning and use of models in mathematics and the empirical sciences*, in The Concept and Role of the Model in Mathematics and Natural and Social Sciences, H. Freudenthal, ed., Dordrecht, 1961, pp. 163–177.

[126] ——, *Models of data*, in Studies in the Methodology and Foundations of Science, Dordrecht-Holland, 1969, pp. 24–35.

[127] ——, *A Probabilistic Theory of Causality*, North Holland, Amsterdam, 1970.

[128] ——, *Set-theoretical structures in science*, tech. rep., Institute for Mathematical Studies in the Social Sciences, Stanford University, 1970.

[129] P. SUPPES AND M. ZANOTTI, *When are probabilistic explanations possible?*, Synthese, 48 (1981), pp. 191–199.

[130] B. VAN FRAASSEN, *A formal approach to the philosophy of science*, in Beyond the Edge of Certainty, R. Colodny, ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1965, pp. 303–66.

[131] ——, *A semantic analysis of quantum logic*, in Contemporary Research in Foundations and Philosophy of Quantum Theory, C. Hooker, ed., Reidel, Dordrecht, 1973, pp. 80–113.

[132] ——, *To save the phenomena*, Journal of Philosophy, 73 (1976), pp. 623–632.

[133] ——, *The Scientific Image*, Oxford University Press, Oxford, 1980.

[134] ——, *Empiricism in the philosophy of science*, in Churchland and Hooker [37], pp. 243–305.

[135] ——, *The semantic approach to scientific theories*, in Nersessian [108], pp. 105–124.

[136] ——, *The Charybdis of realism: Epistemological implications of quantum theory*, in Cushing and McMullin [39], pp. 97–113.

[137] ——, *Laws and Symmetry*, Oxford University Press, 1989.

[138] ——, *Quantum Mechanics: An Empiricist View*, Oxford University Press, Oxford, 1991.

[139] J. WHEELER AND W. ZUREK, eds., *Quantum Theory and Measurement*, Princeton University Press, 1983.

[140] E. WIGNER, *The problem of measurement*, American Journal of Physics, 31 (1963), pp. 6–15. Reprinted in [139].

[141] N. WISE, *Mediations: Enlightment balancing acts, or the technologies of rationalism*, in Horwich [88], pp. 207–258.

[142] R. WOJCICKI, *Set-theoretic representations of empirical phenomena*, Journal of Philosophical Logic, 3 (1974), pp. 337–343.

[143] J. WOODWARD, *Data and phenomena*, Synthese, 79 (1989), pp. 393–472.

[144] J. WORRALL, *An unreal image*, British Journal for the Philosophy of Science, 35 (1984), pp. 65–79.