UNIVERSITÀ DEGLI STUDI “ROMA TRE”

FACOLTÀ DI LETTERE E FILSOFIA

DOTTORATO DI RICERCA IN
FILOSOFIA E TEORIA DELLE SCIENZE UMANE

XIX CICLO

STRUCTURAL EXPLANATION

TUTOR:
MAURO DORATO

DOTTORANDA:
LAURA FELLINE

COORDINATORE:
PAOLO D’ANGELO

ANNO ACCADEMICO 2007-2008

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but "That's funny..."

Isaac Asimov

Introduction

The pervasive role of mathematics in modern science gave life to many concerns in the philosophy of science. Among them a topic of growing interest is the epistemological status of mathematical *explanations* of natural phenomena. In this perspective an extensive literature can be found for instance in cognitive sciences – concerning the so-called *computational explanations* (McCulloch and Pitts 1943, Wiener 1948, Piccinini 2006), where the mental capacities of the brain are explained by its computations – and in more recent times a significant number of papers have been investigating the role of mathematical explanations also in biology (Berger, 1998, Baker, 2005).

Since the role that mathematics plays in the explanation of natural phenomena can hardly be overrated, it seems remarkably odd that such a topic has been hitherto neglected in philosophy of physics, the mathematised science *par excellence*. Far from representing a problem concerning only the philosophy of science, the ignorance of the role that mathematics have in the scientific understanding of physical phenomena is mirrored and

emphasized within the perception that society has of theoretical physics. The high level of abstractness of current physical theories dramatically concurs to the widespread feeling that contemporary physics does not aim at (and anyway fails in) the explanation and understanding of the world, but contents itself with the mere manipulation of symbols for the prediction of phenomena.

The current state of scientific knowledge and within it of the relationship between mathematics and explanation is well illustrated by Ruth Berger:

> "Today's science is often concerned with the behavior of extremely complicated physical systems and with huge data sets that can be organized in many different ways. To deal with this, scientists increasingly rely on mathematical models to process, organize, and generate explanatory information, since much of the understanding produced by contemporary science is gathered during the process of mathematical modelling, it is incumbent upon philosophical accounts of explanation to accommodate modelling explanations. This is recognized by the semantic view of theories, which identifies mathematical modelling as one of the mains explanatory engines of science." (Berger, 1998, p.308)

But to the acknowledgement of the central role of models in science did not correspond the recognition of a similar role in the more restricted field of scientific explanation:

> "Although many philosophers accept the basic features of the semantic view of theories, there have been surprisingly few attempts to reconcile it with our best philosophical accounts of scientific explanation. […] [C]ausal accounts cannot illuminate precisely those explanatory features of science which the semantic view deems most important. Specifically, causal accounts of explanation cannot accommodate, and often obscure, the crucial role which mathematical modelling plays in the production of explanatory information. moreover, evidence from modeling explanations

> indicates that causal relevance is neither a necessary nor a sufficient condition for explanatory relevance" (ibid. p. 308-309)

The need for a deeper investigation on this subject becomes then urgent if we take a look at quantum mechanics, which currently represents the *bête noire* of the theory of scientific explanation. The difficulties that have raised in relating quantum mechanical phenomena to classical concepts like properties, causes, or entities like particles or waves are still open, so that there is not yet agreement on what kind of metaphysics is lying at the foundations of quantum mechanics.

It is for this reason that many philosophers say that they are not ready to take lessons from quantum theory until its interpretation is sorted out, and, in particular, that before we can draw any conclusion towards explanation in quantum theory we have to wait for the interpretational problem to be solved.

On the other hand it is to be considered that the problem of the explanatory power of quantum mechanics seems to be not so pushing, in the main, for the physicists' community as it is for philosophers, and quantum theory seems in the eyes of the former to be as explicative with respect to phenomena as any other physical theory. In spite of the lack of a causal account of quantum phenomena physicists constantly use the formal resources of quantum mechanics in order to explain quantum phenomena, and if on one hand analyzing and, if necessary, questioning the epistemic value or the coherence of some explanations is both a right and a duty of philosophy of science, on the other hand philosophy of science should not dismiss a well-established scientific practice as epistemologically irrelevant because it does not fit with some pre-defined philosophical standards of what scientific explanations in physics ought to be.

Starting from this fact, and contrarily to the method of put forward a definition of a new categorical framework, make it mathematically precise, and then see if it fits well with Hilbert

space—the program of the theorists of structural explanation is to take seriously physicists practice of explaining phenomena with the resources of the formalism alone, and so taking that structure as explicative in itself.

Robert Clifton provides this definition of structural explanation:

"We explain some feature *B* of the physical world by displaying a mathematical model of part of the world and demonstrating that there is a feature *A* of the model that corresponds to *B*, and is not explicit in the definition of the model.

It is natural to call explanations based on this maxim *structural* to emphasize that they need not be underpinned by causal stories and may make essential reference to purely mathematical structures that display the similarities and connections between phenomena." (Clifton, 1998)

The aim of this thesis is therefore to provide a contribution for the development of a full fledged theory of structural explanation. More exactly, on the one hand, I have tried to strengthen the available characterization of structural explanation *vis-à-vis* with some theoretical issues that were rising from its original formulation. On the other hand, I have analyzed some relevant case studies both in order to test the applicability of the elaborated theory, and to propose solutions to those controversial cases whose perception is so different between physicists and philosophers.

The first chapter is dedicated to an introduction to the traditional theories of explanation (deductive nomological, causal, unificationist and pragmatic), and to the illustration of the problems such theories present. This review is necessary given the various references that in the other chapters will be made to these theories. The second chapter is devoted to R.I.G. Hughes' theory of structural explanation. After an exposal of the details of Hughes' theory, I argue that the latter presents two main problems. The first concerns the range of application of structural explanation. It seems that following Hughes structural explanations occur at a "ground level", at the level of foundational theories. On the other hand, the author does not

provide a clear definition of foundational theory. More exactly, a criterion is needed which can warrant the claim that structural explanation applies to special relativity and quantum mechanics but not, say, to the kinetic theory of gases. I argue that Hughes finds such a warrant in Einstein's distinction between principle and constructive theories and in the claim that quantum mechanics belongs, like special relativity, to the first class. This leads to the second problem, the fact that Hughes' structural explanation may presuppose an interpretation of quantum theory as a principle theory, as proposed by Jeffrey Bub—condition which strongly weakens his theory. I then analyze Hughes' attempt of structural explanation of the EPR correlations and argue that it is questionable given its use of controversial extra assumptions that go beyond the given mathematical background of quantum mechanics. In the third chapter I develop more extensively the theory of structural explanation, trying also to avoid the shortcomings of Hughes' version. I stress that understanding the physical phenomena structurally is based on the role that such a counterpart plays in the whole mathematical model. I defend the thesis that there is no predetermined criterion for the applicability of a structural explanation. While it can be said that a structural explanation applies to highly abstract theories and it is sufficient when a causal one is not available, such a criterion does not imply that a phenomenon *either* requires a structural *or* a causal explanation. I also stress that the requirement for a causal rather than a structural explanation is not an objective fact, but is typically determined contextually by the theory of reference and the beliefs and skills of scientists. Furthermore, I test my version of structural explanation within the context of the Special Theory of Relativity—and I discuss a couple of examples of structural explanations in quantum mechanics, involving non-locality and Heisenberg's Uncertainty Principle. The claim that structural explanation exploits the resources of mathematical models displayed by highly abstract theories would obviously be little informative if not supported by a theoretical background about what scientific models are, how they represent their target and what is the relation between

scientific representation and explanation. The fourth chapter is devoted to these themes. I discuss three theories on scientific representation: R.I.G. Hughes' Denotation, Deduction, Interpretation theory (DDI) (Hughes, 1997), Mauricio Suárez' inferential theory (Suárez, 2004) and Gabriele Contessa's interpretational theory (Contessa, 2007), that pivot on two key concepts. First of all if *a* represents *b*, then *a* stands for, refer to *b*. Secondly, the notion already introduced of surrogative reasoning: scientific representation is to allow an informed user to draw inferences about the represented object. The analysis of these theories will be guided by the question of how they manage to account to two semantic '*conundrums*' (Frigg, 2006) that a theory of models must answer: the '*enigma of representation*' and the '*problem of style*'. Besides the case of Hughes' account (which deserves to be treated separately, given that is not meant to constitute a full fledged theory on representation), I argue that the inferential and interpretational theories fail to account for Frigg's semantic problems, and that such failure is due to the lack of an apt account of what Suárez calls the objectivity of scientific models. I therefore define the objectivity of a scientific representation in a determinate inquiry context C as the capacity of a model to support the achievement of the aims (such as explanation or prediction) of C—characterizing in this way objectivity as a three place contextual relation between the model, the target system and the context of inquiry. Finally, I argue that an adequate solution, both to the problem of style and to the enigma of representation, must be grounded on the notion of objectivity as defined above. Between the assets of the proposed view, there is the fact that it well fits both with an antirealist and a realist view of science.

In the fifth chapter I analyze Jeffrey Bub's Quantum Information Theory (QIT) (Bub 2000, 2004, 2005). In particular I try to answer two questions: the first is what kind of explanation of quantum phenomena, if any, does QIT provide. The second, related question is to which extent Bub's parallel between the explanatory capacity of SR and that of QIT is justified.

The last chapter is the fruit of my research on non locality as accounted for in the Many Minds interpretation of quantum mechanics. One crucial motivation for structural explanation is the fact that there is no uncontroversial interpretation of quantum mechanics which satisfactorily deal with the problem of non locality. However, there is in fact a family of interpretations of quantum mechanics, the so called Everettian interpretations, which are widely taken to be completely non local. This assumption has been recently challenged by an argument proposed by Meir Hemmo and Itamar Pitowsky (Hemmo and Pitowsky, 2003) on the nonlocality of the Many Minds Interpretation (MMI, hereafter) in the version given by David Albert and Barry Loewer (Albert and Loewer, 1988, and Albert, 1992). In the last chapter I discuss and criticize Hemmo and Pitowsky's argument.

Chapter 1

Scientific Explanation

In this chapter I propose a quick review of the most important theories on scientific explanation and understanding. This review will be useful in the next chapters, given the continuous reference that will be made of them overall the rest of this thesis.

Sections 1 concern the Deductive Nomological model, which was the first available systematic account of scientific representation. Sections 2 and 3 will be concerned respectively with the Causal Mechanical and the Unificationist theories—these represent the theories which have dominated the most part of the discussions on scientific explanation. Section 3.2 will be then devoted to Gherard Schurz and Karel Lambert's unificationist theory of scientific understanding. Finally, in section 4 I will consider the pragmatic approach. Finally, section 5 will be devoted to Henk de Regt and Dennis Dieks contextual theory on scientific understanding.

1.1 The Deductive Nomological model

The first organic theory of scientific explanation is due to Carl Hempel and Paul Oppenheim (Hempel and Oppenheim, 1948), although its more articulated version was then formulated by Hempel (1965). According to the Deductive Nomological model (DN) scientific explanation is an *argument which shows how the statement expressing the explanandum event can be logically deducted from some set of initial condition by means of general laws*—if among the laws used in the deduction there is at least one statistical law, the explanation is said *Deductive Statistical* and, in this case, in a good explanation the deduction has to attribute a high probability to the explanandum. Probably the most deadly objection that has been advanced against the DN model is that it is unapt to catch the *explanatory relevance of explanans* (Salmon 1984). The classic counterexample showing such shortcoming in the literature was the case of John Jones' which become pregnant by regularly consuming his wife's birth control pills. According to the DN model, the argument that John Jones did not get pregnant by because a) he consumed his wife's birth control pills (initial condition) and b) every man who regularly consumed his wife's birth control pills do not get pregnant is a scientific explanation (universal law). This example shows that the DN model cannot discriminate between relevant and irrelevant explanatory facts. Another fundamental shortcoming of the D-N model was the *high probability requirement*, the consequence of which was that if an event *e* has a low probability and, nonetheless, it happens, the subsumption of *e* under the set *C* of initial circumstances and laws involved in *e* come into being is not, in the D-N model, a good explanation. In fact, following the D-N model, low probability events are always *inexplicable.*

As a final remark notice that, given his deductive structure and the role that laws play within it, the DN model finds his natural locus in the neopositivist conception of scientific

theories, while it have a more controversial application in the context of the semantic view of theories.

1.2 The Causal Mechanical theory of explanation.

1.2.1 introduction

According to the causal theory of explanation, the concept of causation is at the core of scientific explanation and no scientific explanation is possible without a causal account of the processes that bring into being the event-to-be-explained, the *explanandum* event. The main advocate of the causal theory is Wesley Salmon. In criticizing Hempel's DN model, Salmon firstly proposed his *Statistical Relevance* (SR) model, in which explanation is based not in the high probability requirement, but in the *relations of statistical relevance between explanans and explanandum*. Explaining, in this sense, is showing which initial conditions are relevant to the assignment of probability to the explanans, and in which way (the *positive* or *negative* relevance are both considered in the explanation). However, Salmon abandoned also this theory, for he realized that a complete and deep account of explanation cannot just show how we may logically expect the explanandum event, nor can it confine itself to spotting and listing the statistical relevant relations between initial conditions and explanandum event. In his *Scientific Explanation and the Causal Structure of the World* (Salmon 1984) he states that "what constitutes adequate explanation depends crucially upon the mechanisms that operate in our world" (Salmon 1984, p. 240). According to the Causal Mechanical (CM) model, therefore, a scientific explanation consists in the individuation of the patterns that constitutes the structure of the world and in the demonstration of how the explanandum event fits into these discernible patterns. The conception according to which an explanation shows the mechanisms (causal or not) which produce the facts we are trying to explain is called by

Salmon the Ontic conception of scientific explanation. We will see later that it is against this kind of conception of explanation that R.I.G. Hughes will propose his theory of structural explanation. As Salmon put it (Salmon 1984, pp. 240-242), his CM model of explanation does not impose causal explanation as an a priori condition for scientific explanation and he claims that his only request is that "in view of [the causal explanation] success we would be reluctant to relinquish this explanatory principle" (Salmon, 1984, p.240). Salmon position could be summarized in this way: according to his conception of scientific explanation, explaining means showing how things work, and we have good reasons to think that the basic structure of the regularities of the world is a causal structure. Hence, the CM model of explanation does *not* demand a special role for causality because scientific explanation *need to be causal per se*, but because we have good reasons to think that—apart from the quantum mechanics domain, where events are not fitted in a causal chain—*causation is the key of the patterns that constitute the world*.

In this respect Salmon put the basic feature of explanation in the concept of causality and he then gave a definition of causality which underwent several corrections, in order to go through the serious problems carried by the definition of causality.

1.2.2 Two Attempts: the Causal Mechanical View and the Problem of Relevance.

As we have seen in the introduction to this section, one main claim behind the CM view is that science explains by correctly delineating the causes of the explanandum. A first problem, however, is to find a reliable definition of cause. In this section I will review two formulations that Salmon provided. The first was proposed in (Salmon, 1984), the second was proposed in (Salmon, 1994) in order to answer to some objections advanced against the first. The illustration of these accounts will serve to highlight one main problem of the CM theory, that

is the capacity of picking out those factors which are causally (and explanatorily) relevant for the occurrence of the explanandum.

In his Scientific Explanation and the Causal Structure of the World, Salmon identifies as the fundamental entities of causality processes, instead of events. He characterizes events and processes in terms of their geometrical representation in spacetime: "Events are localized in space-time, while processes have much grater temporal duration and in many cases, much grater spatial extent" (Salmon 1984, p.139)

Besides processes, the other element of causality stressed by Salmon is the concept of *casual interaction*, i.e. an intersection between causal processes in which the structure of both of them is altered. In the geometrical representation of spacetime an interaction is represented by an intersection between two life-line.

Now, inside the class of processes, there are genuinely causal processes and what he calls pseudo-processes. Processes cannot go faster than light, because they are capable of serving as signals, while pseudo-processes can. For a clear cut distinction of the two cases he develops then the notion of *mark*, already introduced by Reichenbach, which can intuitively be represented as a signal, like, for instance, a scruff on the surface of a baseball. What distinguishes genuinely causal process to pseudo-processes is the capacity that only the former possess to transmit a mark, i.e. of conserving into time an eventual modification introduced in process:

"Let P be a process that, in the absence of interactions with other processes, would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the space-time points A and B (A≠B). then, a mark, (consisting of a modification of Q into Q'), which has been introduced into process P by means of a single local interaction at point A, is transmitted to point B if P manifests the modification Q' at B and at all stages of the process between A and B without additional interventions." (Salmon, 1984, p. 148)

A genuinely causal process is, thus, in this representation, a process which is capable of transmitting its structure, that is, it is capable of supporting the causal system in which the signal is inserted and of communicate its variations.

The idea that causal processes generates a *causal structure* equipped with a certain stability and persistency it's at the base of the developing of *forks*, an analytic instrument which Salmon inherits from Reichenbach. This analysis is lead in statistical terms in order to provide an account of causation which can get along with indeterminism and in general with statistical laws.

The principle of *conjunctive forks*[1] then permits to causally connect two or more correlated but not interacting processes A and B, by means of a connection of them to a set of common previous conditions C without which the mentioned correlation would have be seemed highly improbable. On the other hand, the conditions requested by the principle of *interactive forks* governs the modality in which physical interactions determine modifications in the causal structure.

The shortcoming met by the above formulation of the Causal Mechanical view turned out to be several. First of all, the principle of mark transmission heavily relies on counterfactuals. Moreover, it fails to solve the problem of relevance seen about the DN model. This problem is evident if we think again to the case of John Jones: the birth pill taken by John Jones should

[1] In a nutshell, a congiuntive fork is a statistical structure elaborated by Hans Reichenbach in his *The Direction of Time* (1956) in his attempt to explicate the *principle of common cause*. The latter states, roughly, that when apparent coincidences occur that are too improbable to be attributed to chance, they can be explained by reference to a common causal antecedent. The conjunctive forks is defined in terms of the following four conditions:

1) $P(A.B \mid C) = P(A \mid C) \cdot P(B \mid C)$
2) $P(A.B \mid \text{non } C) = P(A \mid \text{non } C) \cdot P(B \mid \text{non } C)$
3) $P(A \mid C) > P(A \mid \text{non } C)$
4) $P(B \mid C) > P(B \mid \text{nonC})$

Conditions 1-4 entail:

5) $(A.B) > P(A) \cdot P(B)$ (see Salmon 1984, cap. 6)

undergoes a series of causal processes in his body, transmitting in Salmon sense a mark. However, the birth control pill is obviously irrelevant for the explanation of his not becoming pregnant.

In (Salmon 1994) is therefore proposed another version of causality, which gets explicit inspiration from Philipp Dowe's theory of physical causality and which focuses on the concept of conservation of quantities. This last account is resumed in 3 points:

1. A causal interaction is an intersection of world-lines which involves exchange of a conserved (invariant) quantity.

2. A causal process is a world-line of an object that transmits a non-zero amount of a conserved (invariant) quantity at each moment of its history (each spacetime point if its trajectory).

3. A process transmits a conserved (invariant) quantity from *A* to *B* $(A \neq B)$ if it possesses this quantity at *A* and at *B* without any interactions in the half-open interval *(A, B]* that involve an exchange of that particular conserved (invariant) quantity. (Hitchcock, 1995)

However, Hitchcock (Hitchcock, 1995) demonstrated how even this last definition of causality not only is still problematic (it involves a counterfactual notion), but brings the CM account of explanation back to the same problem of explication irrelevance it was called to solve:

" Suppose that a shadow is cast on a metal plate that has a uniform nonzero charge density on its surface. The shadow then moves across the plate in such a way that the area of the plate in shadow remains constant. The shadow then possesses a constant quantity of electric charge (a quantity that is both conserved and invariant) as it moves across the plate. The shadow is not participating in any causal interaction as it moves; in particular it is not being bombarded with photons as is the spot of light in a similar example discussed by Salmon (1984, 308). By definition 3, the shadow transmits the charge, and by definition 2, it is a causal process. (Hitchcock, 1995, 314-315)

1.2.3 Non Universality of Causal explanation.

Another kind of problem which undermines the reliability of the CM view concerns its generality. We have seen that in theory Salmon does not support an universalistic conception of the causal theory, but that he argues that a scientific explanation must show the mechanism which lies under the occurrence of the explanandum, and given that we have good reasons to think that the basic mechanism of the word are causal, then arguably scientific explanations must be causal explanations. Two genres of objections are in order here. The first comes from the known counterexamples of the relativistic effects explained by special relativity and of nonlocal correlations explained by quantum mechanics, where the explanandum does not have a causal history behind it. The fact that a causal description cannot fit in with our more fundamental physical theories should be a good reason to think that causation *is not* the key of the patterns that constitute the world. Notice that the problem raises from the double nature (statistical, and individual) of causal relations. On the one hand, the statistical analysis shows that the correlated events of measure in an EPR of Bell kind experiment are causally dependent—but that the correlations cannot be accounted for in terms of common causes (see Appendix). This would imply *a fortiori* that the correlations must be produced by a causal process connecting the two events of measure. On the other hand, locality requires that causal processes cannot propagate faster than light—as a consequence it seems that there is no possibility of explaining causally quantum correlations.

But the limited applicability of the CM view does not concern only those cases where there are no recognizable causal processes in course (about them an advocate of the CM view can simply counter that there is no in fact available explanation). In order to support his claim on the central role of the mechanistic world view in science, Salmon makes appeal to the importance of such an approach in the history of science, starting by the XVIII century.

During the second half part of the XVIII and the XIX century mechanic philosophy has played an undeniable crucial part in the development of modern sciences as biology, psychology and social sciences, above all in contrasting the tendencies to teleological and metaphysical explanations. It played, we can say, a prescriptive role, which imposed to reject all those explanations that appealed to principles acting finalistically. In biology, for instance, the mechanical philosophy brought to the dismissal of vitalistic explanations (which wanted the evolution guided by a 'spiritual principle' acting finalistically) with the charge of covering ignorance of facts with a metaphysical term instead of trying to fill this ignorance. However, nowadays biology and the other sciences got themselves independent from mechanics. In particular, the explanations of the behaviour of complex systems and (in biology, psychology, sociology, but in physics as well) are typically non causal. It is an intuitively understandable fact that causal description of the evolution of species, or of the use of cell phones in Italy, or of the origin of AIDS, in terms of the momentum or energy of the particles involved in the process, are useless at the very least and this is because, given the specific and different models and explanations provided by different sciences, the recourse to causal explanation can add no further understanding (Woodward 2003, § 4.3. For the discussion of a typical example of non causal explanations in biology, see Berger, 1998).

1.3 The Unificationist Theory

1.3.1 Traditional Unificationist Theories

The unificationist account of explanation traditionally emphasizes the role of theoretical explanation—it goes back to the fact that scientific explanations are in genre called to explain nature's regularities (which in the unificationist view are called phenomena) and rarely refer to particular events. Its basic idea is that explaining and giving understanding is

fundamentally a matter of reducing the number of laws that rules nature regularities. The first advocate of this view is Michael Friedman. In (Friedman 1974) the foundations of unificationist view are laid down in the following terms: "I claim that this is the crucial property of scientific explanation—science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given" (p. 15). Let's consider the behaviour of gases: their specific-heat capacities, the fact that they obey Boyle's law and Graham's law—all these facts are explained by science with the fact that gases are made of molecules and that these obey to the laws of mechanics. Following the unificationist account of explanation what makes the laws of mechanics explanatory with respect to the Boyle-Charles law, the Graham law and gases specific heat capacities, is the reduction of independent "brute facts" we accept in nature from three to one. However, Friedman's concept of "independent acceptable phenomena" turned to be untenable (Salmon 1990), and his explanation theory, which was based on it, has been abandoned. After Friedman, the unificationist view has been developed mostly by Philip Kitcher (1989). In Kitcher's account, scientific explanation is a derivation of different phenomena from the same pattern: the more different phenomena instantiate the pattern, the greater is its explanatory power.

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate. (Kitcher 1989, p.423)

As for the causal explanation, the importance of unification in the capacity of science of providing understanding of the world is undeniable, and so is Friedman's statement that "A world with fewer independent phenomena is, other things equal, more comprehensible than one with more". However probably not even Kitcher's formulation of the concept of unification manage to grasp the very core of scientific explanation, to account for what makes

scientific explanations *explanatory*. First of all let's take the famous example of a flagpole that casts a shadow: we can construct both an argument that deduces the length of the shadow from the height of the flagpole and one that deduce the height of the flagpole from the length of the shadow. Following the DN model, this should be enough to say that the latter constitute a good explanation and obviously this fact constitutes a shortcoming of the DN model. Kitcher's treatment of this case (Kitcher 1989) goes as fallows: the deduction of the height of the pole from the length of its shadow add no further unification to the account of the height of the flagpole that counts all the facts that brought to its construction—therefore it is not a good explanation; if, then, we decide to substitute the "construction account" with the "shadow length account", we can easily see that we loose in unifying power, because not all the objects in all the contexts cast a shadow from whose length we can deduce their height. That's why the "shadows account" does not constitute a good explanation of the height of the flagpole. In this way Kitcher seems to account for the asymmetry of explanation, but his argument turns out to be unsatisfactory. The discrimination between the "shadow account" and a good explanation, in fact, is not supposed to grounded on the contingent feature of the shadow: "it seems to be part not just of common sense, but of currently accepted physical theory that it would be inappropriate to appeal to facts about the shadows cast by objects to explain their dimensions even in a world in which all objects cast enough shadows that all their dimensions could be recovered" (Woodward 2003, 5.3).

Kitcher's solution the problem of asymmetry of explanation leads us to consider another issue. In scientific practice it is not required that an explanation, in order to be considered acceptable, should provide the most deep and, in Kitcher's view, the most general description of phenomena, and a theory of explanation should mirror this fact. As an example take the explanation of falling objects provided by Newton's law: it is universally considered a good explanation even if a more general account of gravity and of the falling of massive bodies is gave by Einstein's General Relativity Theory. However, during the treatment of the precedent

point we saw that the very feature of being less general than another available account is considered by Kitcher a reason to reject the “shadow account” as a good explanation. Hence, for the sake of the argument, Kitcher’s unificationist view is forced to reject all those kinds of explanations that are less general than others available, as non explicative.

As a final remark, notice that, as with the DN model, the unificationist theory put all the explanatory burden to Laws of Nature. As a consequence, within the context of a semantic view of scientific theories the unificationist theory inherit the issues that concern the relationship between models and Laws of Nature.

1.3.2 Schurz and Lambert’s Unificationist Theory of Understanding

In the reminder of this section I will expose a more recent unificationist theory of scientific understanding, due to Gherard Schurz and Karel Lambert (that provides a very different account of unification and consequently of how unification leads to understanding, and that manage to solve the problem just seen in Kitcher’s formulation.

According to Schurz and Lambert, “to understand a phenomenon P (for a given agent) is to be able to fit P into the cognitive background corpus C (of the agent)” (p. 65). C is called the cognitive corpus of the agent, and contains all statements known or believed by the inquirer. More exactly, understanding is defined as a ternary relation between A, the answer to an *understanding seeking how-question*, a phenomenon P and a *cognitive corpus* C. Schurz and Karel embed therefore their formulation in the logical theory of questions and answers: ‘A sentence A yields understanding of P if and only if it answers the question, "How does P fit into C?"’ (p. 67)

Before seeing how P is to be fitted into C, it is necessary to qualify better C. C contains the cognitive representations of phenomena in the form of statements and can be defined as C as a pair (K, I), where K is a set of phenomena (those believed by the inquirer), and I is a set of inferences (those mastered by the inquirer). An answer A, therefore, in order to make P

understandable must add some new information to C, which can be descriptive, and therefore adding to K (in which case A will be factually or theoretically innovative, depending on if the new information contains unknown facts or laws), or in form of some new inferences, and therefore adding to I, which are not known or not mastered in C and that shows how P is inferable from C.

Now, fitting P into C = (K, I) does not only correspond to find connections between P and other elements of K, to find the place of P in C. It can be in fact that the additional information brought by A in order to establish connections between P and a local subset K' of K lead eventually to some contrast between K' and K, destroying in this way coherence. In order to avoid these kind of situations into the definition of understanding, Schurz and Karel pose the condition:

1) "to fit P into K* means (informally) to *connect P* with parts of K* such that […] the local coherence resulting from the connection of P with parts of K* is not outweighed by a loss of coherence elsewhere in K*." (p. 71)

Applied to scientific understanding, the theory, first of all treats scientific knowledge systems as linguistically represented information systems. The connections between the elements of K are arguments ibs (in the broad sense), i.e. "any pair of a set of statements *Prem* (the premises) and a statement *Con* (the conclusion) where a 'sufficient' amount of information is 'transmitted' from *Prem* to *Con.* This is denoted by *Prem*→*Con.*" (p. 71). According to Schurz and Karel the correctness of a scientific argument cannot, be captured by a sufficient condition, but the authors provide a list of types of arguments which extensionally defines the correctness of a scientific argument. Empirical data are the in this picture the way scientific knowledge grasps reality. As elements of the scientific knowledge, they are also elements of C, therefore all the other elements of C must obviously be coherent also with empirical data. Empirical confirmation, therefore, is a special case of coherence, which the authors calls *coherence with data*. (p. 72)

When it comes then to the connection with scientific explanation, Schurz and Karel's theory leads then to a significant consequence. Schurz and Karel notice in fact that, depending on which theory of explanation is considered, their characterization of understanding clashes with the nowadays widely acknowledged claim that every explanation is an understanding-constituting argument, and vice versa. For what the claim that every explanation must provide understanding is concerned, Schurz and Karel notice that many theories of explanation (like the DN and the CM) are merely local explanation, and therefore not sufficient to account for the global character of explanation implied in (1) (p. 109). The inverse is also problematic in the case of restricting theories as the causal one, since every constituting-understanding argument that does not conform to a causal argument would not be considered explanations. A natural consequence of this argument would therefore be the acceptance of a pluralistic conception of explanation, where different kind of explanation takes place depending on what kinds of elements of C are involved and what kinds of information is required to be added in order to fit P into C.

1.4 The pragmatic view

1.4.1 Van Fraassen: answers to why questions

In the pragmatic theory the MC and unificationist explanations are reduced to special cases of scientific explanation. To attribute a pragmatic feature to explanation means that its features and requirements of acceptance depend on the context where it is found. In this sense we can see how it happens that in determinate contexts causal explanations are successful while in others they are not and how the same happens to unifications.

The pragmatic theory of Bas van Fraassen (van Fraassen, 1980) is not intended to be a theory on scientific explanations, but it also embeds other kinds of explanations. According to

van Fraassen, then explanations are answers to why-questions, and a scientific explanations are those answers which make essential use of scientific knowledge. A question therefore is a proposition expressed by an interrogative sentence; in particular a why-question is a question of the form:

Why (is it the case that) *P*?

where *P* is said the *topic* of the question.

A question is dependent on the context in two ways: by the dependence of the *contrast class* and the dependence of the *relevance relation*. The contrast class is a class $X=\{P_1, P_2,\ldots P_{k\ldots}\}$ of propositions, one of which is the topic of the question and the others are propositions that represent the possible alternatives to the topic. Let's take, for instance, the question

Why did Adam eat the apple?,

If the person uttering this question wants to know why did Adam not *throw* the apple instead of *eating* it, the contrast class will be X={Adam ate the apple, Adam threw the apple}. But someone could also know that Adam ate the apple only for hunger and thus ask why did he ate the *apple* instead of a *good pear*, or instead of another fruit—in this case the elements of the contrast class will be: Adam ate a *pear*, Adam ate a *banana*, Adam ate a *peach*, Adam ate a *prickly pear*, and so on. The contrast class, basic element of a question, is dependent on the context because only by the context we can understand in contrast with which alternatives we have to justify the topic.

Another concept involved in the why-question is the relevance relation, which determines under which aspect we want the topic to be explained. Let's call R the relevance relation involved in a question Q. We say that a proposition A is relevant with respect to Q if A is in relation R with the couple $<P_k, X>$, where P_k is the topic of Q and X is its contrast class. R can be a relation of causation, a religious reason, a logic inference and so forth and even in

this case it is the context that decides which kind of relevance relation is involved by the question.

A great merit of a *pragmatic theory of explanation* is its capacity to plug the successes that the MC theory and the Unificationist theory have in treating some kinds of scientific explanation, into a more universal theory. On the other hand, also van Fraassen theory displays some problems. In (Kitcher and Salmon, 1987) it is shown how the lack of any characterizing condition on the relevance relation leads to the consequence that within the pragmatic theory there is always a way to make any truth proposition A the best explanation for any phenomenon. But probably a more crucial objection against van Fraassen theory that many philosophers nowadays acknowledge is that most likely not all explanations are answers to "why questions". Scientific explanations can be answers to how-questions (Hughes, 1993, p.133-134), or to how-can-possibly-be-that questions, or again to how-did-it-occur-that questions (Salmon, 1989, p.231-232). Even if we then take at face value the assumption that all those questions can be anyway translated into why-questions, sometimes these translations manifestly force the natural meaning of the question. It is legitimate to ask at this point to which extent a theory on answers to why questions may grasp the essence of scientific explanations, whose translation in terms of answers to why-questions sometimes can be made only proviso some loss of meaning.

1.4.2 Achinstein's Illocutionary Theory

Another very important advocate of the pragmatic view is undoubtedly Peter Achinstein (1983). Here I just want to discuss a particular aspect of Achinstein's theory, that is the illocutionary force problem. Achinstein's firstly characterizes the concept of explanation as the product of the illocutionary act which produces it, that is the act of explaining. A theory on scientific explanation, therefore, must before be a theory on the illocutionary explanatory

act, where for our purposes we can be content with Achinstein's minimal condition for an explanatory act:

"S explains q by uttering u only if S utters u with the intention that his utterance of u render q understandable" (Achinstein, p. 16)

Where q Achinstein's view of explanation holds that one "must begin with the concept of an illocutionary act of explaining and characterize explanations by reference to this, rather than conversely" (Achinstein, p.102).[2] Achinstein notices that the same sentence u (for instance, "Bill ate spoiled meat") can be uttered both as an explanation (if the doctor is uttering it to explain the reason of Bill's stomach ache) or as a criticism (if, say, Jane critics Bill because he eat spoiled meat). A necessary condition for the explanatory nature of u, therefore, is not given by any intrinsic feature of u, but by its pragmatic role of standing in the relation "being an explanation of" with a given explanandum q.[3] The difference between an explanation (i.e. the product of an explanatory act) and, say, a description (i.e. the product of a descriptive act) is to be found only in the act with produces it. Consequently, an explanation is described as an ordered couple <x, y>, where x is a determinate kind of proposition[4], and y is an explicatory act.

1.5 The Contextual Theory of Scientific Understanding.

Due to the pragmatic nature that van Fraassen attributes to scientific explanation, he notoriously considered explanation as extraneous to the epistemic dimension of science. In

[2] Holding this view, however, implies neither that for every single utterance of u it corresponds an explanation, nor any ontological stance about explanations as sentences or propositions. For an account of how avoiding these and other possible problems, see Achinstein (1983, § 2, 3).
[3] For the other sufficient conditions that u must satisfy in order to be an explanations, we also refer to Achinstein (1983, § 2).
[4] A "content-giving proposition", see (Achinstein pp. 36-45)

other words, in so far as it is a pragmatic activity, explanation has no epistemic value and, hence, cannot be considered part of the *epistemic aim of science*.

In (De Regt and Dieks, 2004) the above conclusion is challenged, and it is argued that the pragmatic dimension of explanation does not imply its epistemic irrelevance. According de Regt and Dieks the role of unifying concept in the theory of explanation is played by the concept of understanding: explanations explain because they provide understanding and, as a consequence, scientific explanations are explanatory because they provide *scientific understanding*. de Regt and Dieks do not enter in the details of a theory on scientific explanation, their central aim is to define scientific understanding: if explanations give understanding "a general theory of understanding might tell us how" (De Regt and Dieks 2004) and with this purpose De Regt and Dieks introduce two criterions: the Criterion for Understanding Phenomena and the Criterion for the Intelligibility of Theories.

CUP: *a phenomenon P can be understood if a theory T of P exist that is intelligible (and meets the usual logical, methodological and empirical requirements)*.

CIT: *a scientific theory T is intelligible for scientists (in context C) if they can recognise qualitatively characteristic consequences of T without performing exact calculations*.

The *usual logical, methodological and empirical requirements* allow to avoid the use of non-scientific theories to achieve understanding of phenomena, on the other hand, the CIT criterion guarantees the contextual character of understanding: first of all, the definition of scientific intelligibility implies a reference to the scientists who are the subject of the understanding act; second, the insertion of the "in context C" clause restricts the scientists' act of understanding to a determinate context so as to allow the possibility that a theory could result intelligible for a scientist (or a community of scientists) in a certain context and unintelligible for other scientists in another context.

According to De Regt and Dieks, the way scientists can obtain a qualitative understanding of phenomena is, again, a context-dependent fact: qualitative understanding can be achieved

through different *tools* and the choice and the acceptance of such tools is decided by the community the scientists belongs to. In some scientific explanations intelligibility is given by a causal account, in other explanations by reduction to more fundamental laws, in others by visualization.

Chapter 2

R.I.G. Hughes' Structural Explanation

The father and the major supporter of the structural explanation program within quantum mechanics, and the one who provided the most articulated treatment of it is surely R.I.G. Hughes.

A substantial part of the ideas that will be proposed in this thesis are, in a more or less explicit form, already present in Hughes' pivotal works. It is therefore a must to start this essay on structural explanation with an exposure of Hughes' pioneer theory.

Structural explanation is to be seen, in Hughes' work, as a particular case of his general theory of scientific explanation, theoretical explanation, according to which:

> We explain some feature X of the world by displaying a model M of part of the world and demonstrating that there is a feature Y of the model that corresponds to X, and is not explicit in the definition of M. (Hughes, 1993, p.133)

While the natural locus for the DN model of explanation is the axiomatic conception of scientific theories, where theories are conceived as a partially interpreted axiomatic calculi, structural explanation, where models play the central role, is naturally fitted into a semantic view of theories.

2.1 Ideology and Explanation

As we have seen in the introduction, the problem of explanation in quantum mechanics is a derivation of the problem of interpretation. Hughes (1989a and 1989b) illustrates this situation recurring to the concept of *categorial framework*, which is the set of fundamental metaphysical assumptions about what sorts of entities and what sorts of processes lie within a theory's domain (Körner, 1969, pp.192-210). The inexplicability of the EPR correlations, therefore, consists in the fact that the behaviour of a pair of particles in EPR state seems to violate the *categorial framework* at the base of classical mechanics, and "it is not clear what categorial elements we can hope to find represented within [the models provided by Hilbert spaces], nor, when we find them, to what extent the quiddities of these representations will impel us to modify the categorial framework within which these elements are organized"(Hughes, 1989 b, p.176).

If this is true, the problem of the inexplicability of some quantum phenomena seems so to be well diagnosed by Duhem (1954), for which scientific explanations are attempts to account for phenomenal laws in terms of prior metaphysical assumptions.

> If an appeal is made in the course of the explanation of a physical phenomenon to some law which that metaphysics is powerless to justify, then no explanation will be forthcoming, and physical theory will have failed in its aim. (Duhem, 1954)

Such metaphysical assumptions determines, often unconsciously, what is regarded as *natural*, i.e. as in need of no explanation. The set of such assumption is called by Hughes an *ideology*. As a consequence, although often not rationally justifiable, an ideology also lies at the base of what a group of persons takes to be a scientific explanation. The above analysis of the disaccord between philosophers and physicists about the explanatory power of quantum mechanics could be therefore translated in Hughes' terms as a disaccord between the respective ideologies of the two groups, i.e., as a disaccord between what the two groups regard as what is natural and what is not, and what can be considered explained, what not.

The clearest illustration of how this disaccord manifests itself, is the situation depicted in chapter 1 of the causal explanation of the EPR correlations. As it is there explained, under the request of a purely probabilistic account of causality, a causal explanation of the correlations is easily producible. The problem arises instead when, as in Salmon's ontic conception, an account of the causal *processes* involved (and therefore of the kind of entities in such processes involved) is necessarily requested.

When ontological assumptions are considered crucial for an explanation, the explanatory power of quantum mechanics crashes onto its foundational problems.

A way of going beyond this impasse would be to provide an account of the way quantum mechanics explains phenomena which could avoid any metaphysical assumption. But is there any explanation like this?

A first attempt is represented by the DN model of explanation. This is a non metaphysical explanation since it limits to logically deduce the explanandum from initial conditions and laws of nature. No implicit metaphysical assumptions are needed. The problem, as noticed by Hughes, is however that it is not clear what in quantum mechanics could count as the right law of nature explaining the correlations. If we take the very same statements of the correlations as the expression of a law of nature, the correlations would

explain themselves. This, however, would correspond to a simple stipulation, which cannot count as an explanation. On the other hand, according to Hughes, taking the known theoretical principles of quantum mechanics as the required laws, would not clarify much, since this "proffered "explanation" will merely rehearse the quantum-theoretic derivation of the statements of correlation." (1989a, p.197). Notice that it is not clear here what Hughes' criticism to the latter possibility exactly is. Hughes never spells this criticism out. On the other hand, to clarify this point is important since structural explanation is supposed to solve the problems embedded in the DN explanation. We will come back to the relation between DN and structural explanation in the next chapter, and we will see how it plays an important role in the general characterization of structural explanation.

Back to ideology and explanation, it seems that the problem of explanation in quantum mechanics leads to the following dilemma: either we rely on some ideology (and then the problem is to find a categorial framework which fits with the quantum mechanical formalism) or we content ourselves with a pure derivation, which however does not seem to really improve our insight of phenomena.

2.2 Structural Explanation.

Structural explanation, says Hughes, provides us with an alternative: "a structural explanation displays the elements of the models the theory uses and shows how they fit together. More picturesquely, it disassembles the black box, shows the working parts, and puts it together again. "Brute facts" about the theory are explained by showing their connections with other facts, possibly less brutish." (Hughes, 1989a, p.198).

But if structural explanation is a particular case of the more general theory of theoretical explanation, how does the former differentiates from any other scientific explanation?

First of all, Hughes suggests that the peculiarity of structural explanation consists in that it works with the *mathematical* models provided by theories (1989b, pp.255-256).

The typical example of how such an explanation works can be found in the STR. Suppose that we were asked to explain why, according to the STR, there is one velocity which is invariant across all inertial frames[5].

> "A structural explanation of the invariance would display the models of space-time that STR uses, and the admissible coordinate systems for space-time that STR allows; it would then show that there were pairs of events, e_1, e_2, such that, under all admissible transformations of coordinates, their spatial separation X bore a constant ratio to their temporal separation T, and hence that the velocity X/T of anything moving from e_1 to e_2 would be the same in all coordinate systems. It would also show that only when this ratio had a particular value (call it "c") was it invariant under these transformations." (ibid. p.256-257)

On the other hand, according to Hughes, structural explanation occurs in a specific kind of theory:

"Explanation comes at many levels, as does scientific theorizing. It is the foundational level which concerns us here, since it is at this level that structural explanation occurs." (ibid., p.257)

However, it is not evident what Hughes' intend with the expression 'foundational level'. This leaves open the question of when structural explanations occur in physics. In other words: to which cases does the theory of structural explanation apply? And why, then, to quantum theory but not, say, to the kinetic theory of gases?

One way to address this problem is by looking at the only uncontroversial (at least following Hughes[6]) example of structural explanation we have, i.e. the one illustrated above of STR. The above questions, therefore, could be reformulated as follows: on what grounds

[5] The fact that this happens to be the speed of light is, according to Hughes', irrelevant, given that he takes SR as a theory about space-time.
[6] For a detailed discussion of structural explanation in STR see (Dorato 2007) and Chapter 3 of this thesis.

STR's and QM' explanations can be made equal? Where is the possible link between quantum theory and SR? Again, we cannot simply say that the peculiarity of such explanations is that they use the formal resources of the theory, since every physical theory displays some mathematical model, but only *ground-level* (Hughes' term) theories provide structural explanation.

A first hint could then be found in Hughes' classification of scientific models. Following Hughes, in fact, we can distinguish three kinds of models[7]: *constitutive*, *foundational* and *structural*. They differentiates by the generality of their subject (i.e. the part of the world that the theory deals with) and by their degree of abstraction. Constitutive models are models of singulars, as can be Bohr's model of hydrogen atom. Foundational models are more general class of systems, like the systems governed by Newtonian mechanics. Finally, "Newtonian mechanics, presupposes a particular account of space-time. […] Though the distinction between this kind of model and a foundational model is not a sharp one, I call these models *structural models*" (Hughes, 1993, pp. 137-138).

However, as recognized by Hughes himself, the distinction between structural and foundational models is not so sharp as one may wish. While it seems reasonable to put Minkowski's spacetime between the *ground-level* structural models—the same is not so obvious about the models displayed by quantum mechanics, i.e. Hilbert spaces. At the end of the days, the subject of quantum mechanics (quantum *systems*) seems to be more similar to the subject of Newtonian mechanics (Newtonian *systems*)—it is then logical to put Hilbert spaces at the less fundamental and less general level of foundational models, as it is the case of Newtonian models.

[7] As we will see in §2 of Chapter 4 on models and representation, Hughes proposes a different taxonomy for models in each of his works on models and explanation, and not always they are in accord with each other. Here we will propose the one he used for the illustration of theoretical explanation (Hughes, 1993) granted that, basically, the criterion of such a classification is always the same (i.e. the degree of generality of the subject of the model) and that the other taxonomies could fit with his theory of explanation as well as this.

2.3 Principle Theories.

We will now argue that the justification for Hughes' claim that quantum mechanics belongs to the same ground level as STR and that they provide the same kind of explanation is to be found in another, sharper classification with respect to that, more blurred, based on the subject of theories. This is the distinction between principle and constructive theories.

The distinction between principle and constructive theories was originally drawn by Einstein in 1919.

> "We can distinguish various kinds of theories in physics. Most of them are constructive. They attempt to build up a picture of the more complex phenomena out of the material of a relatively simple formal scheme from which they start out. Thus the kinetic theory of gases seeks to reduce mechanical, thermal, and diffusional processes to movements of molecules— i.e., to build them up out of the hypothesis of molecular motion. When we say that we have succeeded in understanding a group of natural processes, we invariably mean that a constructive theory has been found which covers the processes in question.
>
> Along with this most important class of theories there exists a second, which I will call 'principle theories.' These employ the analytic, not the synthetic, method. The elements which form their basis and starting-point are not hypothetically constructed but empirically discovered ones, general characteristics of natural processes, principles that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy. Thus the science of thermodynamics seeks by analytical means to deduce necessary conditions,

> which separate events have to satisfy, from the universally experienced fact that perpetual motion is impossible." (Einstein, 1954)

Jeffrey Bub has then drawn on this distinction and argued that QM, as STR, is a principle theory[8].

Hughes endorses the parallel that Bub draws between quantum mechanics and STR, and Bub's interpretation of quantum theory "as a theory which postulates novel, non classical "possibility structures" (Bub's phrase)" (Hughes, 1989a, p.205).

This is a detail of Hughes' conception of structural explanation that we think is worth stressing. According to Hughes, principle theories provide structural explanations:

> "Whenever we appeal to a principle theory to provide a theoretical explanation, I claim, the explanation consists in making explicit the structural features of the models the theory employs. In the same way that we explain the constancy of one particular velocity with respect to all inertial frames by appealing to the structure of Minkowski space-time, we explain paradoxical quantum-mechanical effects by showing, first of all, how Hilbert spaces provide natural models for probabilistic theories […], and, second, what the consequences of accepting these models are" (1989b, p.258)

It is evident how such an interpretation is meant to solve the problem of the pervasiveness of ideology in scientific explanations. For its very nature, in fact, a principle theory expressly dismisses any talk of a categorical framework of reference (would it be made up of an ontology of non-classical waves or particles), and limits itself to supply "models which display the structure of a set of events" (Hughes, 1989b, p.258).

[8] Hughes relies to Bub (1979) and (1974). But see also Chapter 4 of this thesis and relative bibliography for a more detailed treatment of Bub's theory.

To be sure, there is no place in Hughes where it is suggested that *only* principle theories provide structural explanations, however, the existence of an intimate link between principle theories and structural explanation is undeniable, mostly when it comes to quantum mechanics. In other words, in the reading we are suggesting, it is from quantum theory's nature of principle theory that Hughes finds justification for its comparison between SR's and quantum mechanical explanations.

Two problems are implied by this assumption. The first is the obvious fact that any account of structural explanation in quantum mechanics can be taken at face value only provided that also the principle interpretation of quantum mechanics (viz. Bub's theory) is. The second is that under such a condition, we arguably loose an important asset that we have claimed about the application of structural explanation to quantum mechanics, i.e. its capacity of accounting for the attitude of working physicists towards explanation in quantum mechanics. It is reasonable to assume, in fact, that working physicists does not take quantum mechanics as a principle theory as STR, not even in a somewhat "unconscious" or implicit way.

This is, we think, a good reason to search for another account of the range of application of structural explanation in physics, one which relies on a distinction that could be more reasonably related to physicists work

2.4 Structural Explanation of the EPR correlations.

IL PROBLEMA DI HUGHES è CHE CON QUESTO ESEMPIO SPIEGA LA STRUTTURA DI UNA PROBABILITà. E SE VUOLE RIMANERE Lì VA BENE. MA SE QUELLO CHE VUOLE SPIEGARE SONO I FENOMENI, ALLORA DEVE DARE UNA GIUSTIFICAZIONE perché QUESTA STRUTTURA DELLA PROBABILITà CI DOVREBBE SPIEGARE I FENOMENI. MA QUESTO, DI NUOVO, IMPLICA

PROBLEMI. SINO A CHE PARLA DI PROBABILITà IL PROBLEMA DELLA MISURA NON C'è, MA NON è ARRIVATO A SPIEGARE LE CORRELAZIONI, MA QUANZO POI SI CERCA DI PASSARE DALLA PROBABILITò ALLE CORRELAZIONI, ALLORA C'è IL PROBLEMA DELLA MISURA A practical example of structural explanation is provided by Hughes' in the explanation of the EPR correlations (1989a).

Quantum mechanics, says Hughes, provides us with a model, Hilbert spaces, of the observables of quantum systems. This model has the following features:

I) Representability: quantum system's states and the observable physical quantities are to be represented within Hilbert spaces. An important consequence of Representability is that there are families of observables (say, the family of components of spin of a particle) which are systematically related to each other, in such a way that a pure state of a system assigns probability 1 only to a value of only one observable of the family.

II) Non-orthodoxy: the probability assignments of quantum theory do not conform to the standard (orthodox) mathematical account of probability, due to Kolmogorov. This is due to the fact that the orthodox account of probability defines it as an additive measure on a field of subsets on a given set with a Boolean structure. Classical theories' probability functions conform to this definition because the set of propositions of a classical theory, represented by subsets of a phase space, have a Boolean structure. On the other hand, the propositions of quantum mechanics are represented by subsets of a Hilbert space, whose structure is an ortho-algebra, a set of Boolean algebras pasted together in a consistent way.

III) Projectability: as a consequence of non-orthodoxy, the conditionalized probability of a measurement outcome (P) on another (Q) for a pure state d is equal to the probability that the projection d' of d on Q provides for P (Lüders Rule). Projectability is the thesis that conditionalization *does in effect project d to d'*.

IV) Entanglement: the state of two coupled quantum systems *a* and *b* is represented by the tensor product of the Hilbert spaces associated with *a* and *b* with the following 4 features:

i) Each pure state *W* of *C* can be uniquely "decomposed" into "reduced" states, w_a, w_b of *a* and *b*.

ii) Even when *W* is a pure states, w_a, w_b may be "mixed states, that is, states which behave like weighted sums of pure states.

iii) The cocnverse of (i) does not hold;when w_a, w_b are mixed states, there can be more than one state *W* which reduces to them.

iv) There are states *W* such that *a* and *b* are not statistically independent of one another. (p.201)

Now, let's take an EPR-Bohm experiment. With two spin ½ particles *a* and *b*, in the singlet state: $W = \frac{1}{\sqrt{2}}\left(|+\rangle_a|-\rangle_b - |-\rangle_a|+\rangle_b\right)$

For Representability, the state *W* of the coupled system can be represented in a Hilbert space.

For Entanglement [i] and [ii], even *W* is a pure state, the states $w_a = \frac{1}{2}\left(|\uparrow\rangle_a\langle\uparrow|_a + |\downarrow\rangle_a\langle\downarrow|_a\right)$ and $w_b = \frac{1}{2}\left(|\uparrow\rangle_b\langle\uparrow|_b + |\downarrow\rangle_b\langle\downarrow|_b\right)$ of *a* and *b* are improper mixtures. For both *a* and *b*, the initial probability of obtaining the + value in a measurement of any component S_i of spin is ½.

We must explain why, if we perform a measurement of, say, S_x, on *a*, and it yields +, the probabilities are as though system *b* were in the eingenstate which gives probability 1 to the event (S_x^b,-) of obtaining a – result to an S_x measurement on *b*. Notice that, as observed by Hughes himself, this result is already contained in Entanglement plus Representability, given that for Entanglement (*iv*) a and b are not statistically independent.

To understand the EPR correlation, therefore, following Hughes, we need Projectability. Once we apply Lüders Rule, in fact, we see that the conditionalization on the event (S_x^a, +) project *W* into the state $W' = x_+^a \otimes x_-^b$.

Now, according to Hughes, the above account is not yet a structural explanation of the EPR correlations: "I can imagine the objection being made that as long as the alleged explanation stays within quantum theory it is uninformative, and that where it goes beyond the theory, it is suspect" (p. 203). To provide a structural explanation, Hughes then intend to justify the (I), (II), (III) and (IV), in such a way to still avoid any recourse to ideology.

The first two principles are purely formal principles, and do not need any special justification: "in enunciating Representability and Entanglement, I am doing no more than drawing attention to the models that quantum theory employs" (p.203)

The same is not valid for what Non-Orthodoxy and Projectability are concerned, for the simple reason that they are not strictly formal principles, i.e., elements or features of the Hilbert formalism.

Notice that Hughes switches here the level of the discussion. In particular, with Non-Orthodoxy he passes from a description of the problem in terms of Hilbert spaces, to a description in terms of probability structures defined over such spaces. Intern to such a description, Projectability is then easily justified as the quantum analogue of orthodox conditionalization, and therefore a natural corollary of Non-Orthodoxy.

As a conclusion:

> "Representability and Entanglement account for these correlations by giving the quantum theoretic analysis of them. Non-Orthodoxy is an invitation to look at a particular aspect of this analysis, and to see how the results derive from the nonclassical nature of the probability functions involved. One could reasonably claim that this alone would constitute a structural explanation of them. However, we can reinforce this invitation by

> pointing out that the probabilities involved in the EPR experiments—including those that yield the correlation—are often conditional probabilities"

and that as such they are governed by Lüders Rule.

We do not want to discuss here the legitimacy or not of the assumptions of Non-Orthodoxy and Projectability. As Hughes itself is well aware of, its arguments can demonstrate that they are reasonable at best, but not that they are a straight consequence of Representability and Entanglement. But this is not the point here.

Hughes' introduce the meaning of Non-Orthodoxy and Projectability by saying that "the provision of structural explanation consists in *making explicit* what is involved in [subscribing to Representability and Entanglement]. The first move in this direction is made with the thesis of Non-Orthodoxy" (ibid., my emphasis). However, such "making explicit" is something very different from the declared means of structural explanation, i.e. to "display the elements of the (mathematical) models the theory uses and shows how they fit together", to display a model M of part of the world and demonstrating that there is a feature Y of the model that corresponds to the explanandum. For as neutrally as Hughes tries to present the description of quantum mechanics "as a theory which postulates novel, nonclassical "possibility structures"", it still remains the fact that with this description we leave the neutral field of mathematical models and enter to all intent and purposes to the field of the interpretation of quantum theory!

The point we want to make is that there is a long way from saying that Non-Orthodoxy and Projectability are legitimate to saying that quantum mechanical phenomena can be accounted for in terms of such assumptions. Just to start, once pure mathematics is left and an interpretation in terms of structure of probabilities is adopted, the question raises naturally: probability of what? Of the result of the experiments? But is it then an instrumentalist view that Hughes' is endorsing here? More in general, if, as it is legitimate to

think, Hughes is also in this case relying to Bub's interpretation of quantum theory, then the consistency of Hughes' account hinges upon the acceptance of Bub's program.

Chapter 3

Structural Explanation, again.[9]

3.1 Introduction

It will be useful at this point to recap the conclusions of our analysis of Hughes' theory.

A first problem to be addressed is that the original formulation provided by Hughes and further characterized by Clifton of structural explanation, is, as it stands, manifestly incomplete. The first urgent problem concerns the range of application of structural explanation. The Hughes' approach to the problem, relying on a distinction between theories on the basis of the generality of their subject, or on the Einstenian distinction between principle and constructive theories, is shown to be unsatisfactory. In this chapter I will argue that there is no predetermined non contextual criterion for the applicability of a structural explanation. While it can be said quite vaguely that a structural explanation applies to highly abstract theories and it is sufficient when a causal/mechanical one is not available, such a

[9] A lot of the conceptual groundwork for this Chapter was laid in Dorato and Felline (2008). I am therefore highly indebted to my coauthor.

criterion is not meant to imply that a phenomenon *either* requires a structural *or* a causal explanation[10]. Instead, a view were the two different explanations of the same phenomenon can also cohabit in science will be advocated. Clearly such an approach requires an analysis of the consequences of the pluralistic view to which it is committed—more in particular, the relation that in such a view is assumed to hold between causal and structural explanations, needs to be spelled out more in details.

Another open problem concerns the distinction between structural explanations and pure formal derivations of a physical phenomenon. I will argue in favour of a pragmatic view of explanation, and that, therefore, the question is to be seen as a part of the more general problem of the discrimination between explanations and simple descriptions which Achinstein called the *illocutionary force problem* (Achinstein 1983, §3). Following Achinstein I will maintain that what firstly distinguishes a structural explanation *u* for any mathematical account is given by its pragmatic role of standing in the relation "being an explanation of" with a given explanandum *q*.

I will then argue, borrowing from de Regt and Dieks' contextual theory of understanding, that an important role to contextual factors as the skills and beliefs of the scientists for the evaluation of an explanation (de Regt and Dieks, 2003).

The discussion carried in the following will also display another difference with respect to Hughes' work, and which is due to the main background motivation that lead the overall research of this thesis. I have argued in fact that a main asset of structural explanation is that it offer a new strategy for the solution to those controversial cases whose perception is so different between physicists and philosophers.

Contrarily to the traditional attempts to account for the explanatory power of theoretic physics is that its strategy is program is to take seriously physicists practice of explaining

[10] In the following, for short I will indicate the causal/mechanical explanation with 'causal explanation'. With causal explanation we will therefore intend any explanation that shows the mechanisms (causal or not) which produce the facts we are trying to explain

phenomena, rather than dismiss a well-established scientific practice as epistemologically irrelevant simply because it does not fit with some pre-defined philosophical standards of what scientific explanations in physics ought to be. As a consequence, the case studies which will be presented in the present chapter will be aimed to show how structural explanations are *already* used and commonly accepted among physicists.

Although the typical example of structural explanation in physics is indubitably Special Relativity's explanation of relativistic effects, this chapter will be mainly concerned with quantum mechanics. The first case study concerns Heisenberg's Uncertainty Relations. I will argue that the existence of a minimum for the product of the uncertainties of these two measurements is typically explained by means of a structural explanation, i.e. by showing that, given the properties of the formal model used by the theory (the Hilbert space of square summable functions), the formal representative ($\Delta x \bullet \Delta p \geq \frac{\hbar}{2}$) of Heisenberg's relation, is part and parcel of the formal model.

The second case study concerns non-local quantum correlations. These are the infamous 'unexplainable' phenomena which represent such a conundrum for philosophers. I have argued that Hughes (1989a) account of the non-local EPR correlations is questionable given its use of controversial extra assumptions that go beyond the given mathematical background of quantum mechanics. In the context of the present program it is proposed instead to account for the 'working physicist' unproblematic perception of non-local correlation, by exploiting the sole formal apparatus displayed by quantum mechanics.

3.2 Structural Understanding.

By appropriating with some modifications R.I.G. Hughes' 1993 definition of scientific explanation, Robert Clifton writes:

«We explain some feature *B* of the physical world by displaying a mathematical model of part of the world and demonstrating that there is a feature *A* of the model that corresponds to *B*, and is not explicit in the definition of the model.» (Clifton 1998, p.7)

Of course, the peculiar kind of understanding that structural explanations offer needs to be further qualified. The second feature of a structural explanation quoted above −"that *A* is not explicit in the definition of the model"− provides a first hint as to how to tackle this task. Clifton argues that the above requirement is meant to avoid cases of spurious unifications by mere cataloguing the phenomena to be explained—in doing this he follows Kitcher's unificationist account following which to explain means to reduce the number of laws covering phenomena. However, the spurious unifications is notably a problem in Kitcher's view, which lead Clifton's suggestion also controversial.

In the previous chapter it has been illustrated how following Hughes

> "a structural explanation displays the elements of the models the theory uses and shows how they fit together. More picturesquely, it disassembles the black box, shows the working parts, and puts it together again. "Brute facts" about the theory are explained by showing their connections with other facts, possibly less brutish." (Hughes, 1989a, p.198)

The idea here is that it is the very process of *making explicit* the function of *A* in the formal model that contributes to our understanding the physical explanandum *B*. What is needed, in other words, is a clarification of the nature of the connections between the representative *A* of the physical *explanandum* (*B*) and the other elements of the model, which is obtained by articulating the role of *A* in the mathematical model, that is, by making explicit its function. The act of "fitting the explanandum" in the model typically provides understanding by showing the role that *A* has in the whole structure, a role that depends on

A's structural features, and that might have not been clear or explicit when the model was first proposed.

I have argued before against Clifton's reference to Kitcher's unificationist theory. Instead, the proposed account of how structural explanation yields understanding displays many points of convergence with Schurz and Lambert's unificationist theory of understanding. As in Schurz and Lambert's theory the process of understanding P involves to be able to fit P into C (the cognitive corpus), in the same way understanding structurally a phenomenon P implies the capacity of fitting P's representative into the model M. Remember that in Schurz and Lambert's theory, it is admitted that the non understanding of P is not due to the lack of any descriptive information, but can be due to the ignorance of or the inability to master some new inference. In these cases the additional information required in order to make P understandable contains no new facts or law, but consists in the display of the inferences of P from some premises X already known in C. This is typically the case of structural explanation, where all the elements necessary for "putting P into M" are already present in M, and what is needed is an inference that shows how P is connected to these elements.

The process of understanding a physical phenomenon structurally therefore involves reflecting on the defining properties of its formal counterpart, as well as locating such a counterpart in the structure provided by the model. But this also means (perhaps unsurprisingly) that the understanding that structural explanation provides of the explanandum is typically a relational understanding, as the one typically provided by manuals of mathematics when they provide implicit definitions of the mathematical structures and objects (Shapiro, 2000, p.283) See also (Shapiro 1983 and 1997).

3.3 Applicability

In the previous chapter we have seen how a crucial unresolved problem in Hughes' account is that it lacks of a satisfactory account of the range of applicability of structural explanation. More exactly, no explanation is provided for the reason why structural explanation should be applied both to quantum mechanics and to special relativity, and not, say, to the kinetic theory of gases. In this section I want to propose a different approach to the problem. Well, in fact we do not think so. It is our physical knowledge in its globality that can tell us what needs to be explained causally and what not, and errors cannot be prevented. Think of those historical cases in which a phenomenon that, in a certain historical context, was thought to require a causal explanation in terms of a power or a *force* (respectively, the continuation on the part of bodies in a certain state of motion before Galileo, free fall before Einstein), after the revolution is not in need of any explanation (the continuation of motion is "natural), or needs only a structural explanation.

And not only is there no objective way to individuate when an explanandum needs a structural or a causal explanation, but sometimes two explanations of the same phenomena can be acceptable. Maybe the best illustrating example for this is provided by Wesley Salmon. Salmon argues that there are two different explanations for the fact that a helium-filled balloon in an accelerating plane move forward: the first, structural, is provided by general relativity with the principle of equivalence[11], and the second, causal, shows how the acceleration induces a pressure in the back of the plane, and this, in turn, pushes the balloon from the back to the front (Salmon 1998, 73). Notwithstanding the deep difference between the two accounts, both explanations are illuminating from their own different perspective.

Notice that the claim that different explanations can cohabit in science is not new at all. Given the various failed attempts of formulating a theory on scientific explanation in terms of

[11] We follow Salmon in taking this as a structural explanation, however the principle of equivalence is not exactly a formal, purely mathematical element, so the structural nature of the explanation could be questioned.

a unique explanatory model, this idea is nowadays more and more accepted. In fact, such an idea is at the core of the Contextual Theory on Scientific Understanding developed by Henk de Regt and Dennis Dieks (see de Regt and Dieks, 2005). Under the light of such a general theory of understanding, it becomes quite natural to claim that a phenomenon *A* might be regarded as in need of a causal explanation in the epistemic context *x* and can be regarded as explained structurally in the epistemic context *y*, or even that in the same context it may receive two different explanations.

3.4 Evaluation.

The choice of the approach for the solution of the problem of the application of structural explanation has then important consequences also in the relationship between structural and causal explanations. According to Clifton, the first characterization of a structural explanation is a "negative" one, namely that «it need not be underpinned by causal stories». We take this to mean that structural explanations do not necessarily rule out the existence of *causal* stories, nor are necessarily opposed to "physical" or causal types of explanations, or to that effect, to causal theories of explanation in general. In principle, however, structural explanations offer a way of *understanding* physical phenomena that is *independent* of locating the explanandum in the causal fabric of the world. In other words, structural explanations can be useful in order to show that even in the absence of causal or dynamical stories, physical theories can offer in any case a way of understanding physical phenomena, something allowing us to claim that physics is not reduced to a mere quantitative description of phenomena.

I think that Hughes would agree with the last claims, however, it is not completely clear what according to him is the exact relationship between structural and causal explanation. In

order to see why, consider the following possible argument: structural explanation seems to be entirely appropriate in the context of the STR, since the latter is a theory about a particular structure, that of *space-time*. However, the example of STR seems to be of no use for a theory like quantum mechanics which, while obeying a non-classical probability structure, it is not *about* such a structure, but about physical entities and physical quantities, a reference to which is therefore indispensable for a fully *physical* understanding of phenomena.

Hughes' answer to this objection is that, given that the elements of what he refers to as "the categorial framework of reference" (the postulated ontology of the theory) will have to obey the constraints posed by the mathematical structures used by the theory, «it is not clear what useful explanatory work this interpretation would perform over and above that provided by a full articulation of the models the theory presents» (Hughes, 1989a, p.207).

Given the discussion in (§???) it should be recognizable how this quotation is embedded in the context of Hughes' view of quantum mechanics as a principle theory. Consequently, according to Hughes no interpretation of the quantum formalism could help to *account* for the founding principles of the theory, «rather, it would *restate* them in another vocabulary». (Hughes 1989a, p. 207). In Bub's own words, measurement processes are meant to be "black boxes", that quantum mechanics cannot further explore, so that «a principle theory is the best one can hope to achieve as an explanatory account of quantum phenomena» (Bub, 2005, p.558).

I have already argued that to condition our account of structural explanation to a 'principle interpretation' would eventually strongly weaken the likeness of the theory. Moreover, if structural explanation occurs in principle theories, the above argument should be valid also for thermodynamics, while it is obvious that in this case the causal explanation of thermodynamic phenomena in terms of the kinetic theory of gases leads to an undeniable gain of understanding. This last observation lead us then to the main point of this section. Besides its reliance to a principle/constructive distinction, Hughes' argument also shows a

questionable “absolutistic” conception of explanation. In other words Hughes’ argument ends up to exclude or block the possibility of a further deeper explanation of quantum phenomena. Such a stronger conclusion, however, is not only not needed, but undesirable, due its implications in a more general conception of scientific explanation. As in the case of the STR, also in quantum mechanics the adequacy of structural explanations *per se* need not exclude other forms of explanations, and its importance in the quantum domain lies in the fact that obscurantist arguments of the kind “we will never understand the quantum world because its scale is too remote from the mesoscopic one in which we evolved” would be blocked. What we reject is the view that the availability of a structural explanation implies the exclusion of a future causal account. Not only is this not a reasonable request with respect to structural explanations, but towards explanations in general. In a word, the request that a good scientific explanation exclude the possibility of another explanation agrees with what Popper called the doctrine of *essentialism*, which asserts that the best theories can catch the “essential nature” of things and provide us with “ultimate explanations”, which are not further explainable.

Just to give an example, to explain the fact that we have determinate outcomes in a quantum experiment by adducing the fact that the wave function collapses due to some gravitational effect would be *more enlightening* than merely showing how the former fact corresponds to Born Rule within the formal model − provided, of course, that we had some *independent evidence* for accepting the link between wave collapse and gravity, over and above its explanatory power.

To be sure, there are cases in which a structural explanation provides us with a new picture of facts where the search for a causal mechanism bringing about the explanandum seems to dissolve. As we argued before, this is the case of the relativistic effects, and could be the case also for the existence of entangled states.

In a word, the importance of a physical interpretation of the formalism for our understanding of quantum phenomena does not deprive structural explanations of their

potential to make us understand the physical world. True: any interpretation of quantum mechanics increases our understanding of the theory, even if it is underdetermined by the currently available data, but structural explanations, being common to all such interpretations insofar as they share the same mathematical model, provide in any case a unifying explanation of the quantum phenomena, independently of any particular interpretation.

The discussion in the present and in the previous sections surely reveals that I'm proposing a pluralist account of scientific explanation in general. A causal and a structural explanations of the same phenomena can coherently cohabit in science. But can the same be said of two different structural explanations? The fact that with structural explanation the explanandum is understood relationally should not imply a choice when different mathematical models are available for the same theory? This for instance is exactly the case of quantum mechanics, with the standard vector space formalism and the algebraic approach. However, I don't think that the existence of two different formalism of the same theory should necessarily represent an embarrassment for someone advocating structural explanation. First of all, a system can undergo different structural descriptions (without contradiction), and depending on how a system is described, it can exhibit different non isomorphic structures:

"If a system is to have a structure it has to be made up of individuals and relations. But the physical

world does not come sliced up with the pieces bearing labels saying 'this is an individual' or 'this is a relation '. What we recognise as individuals and what relations hold between these depends, in part at least, on what scheme we employ for 'cutting up' the system. But different schemes may result in different structures. So there is no such thing as the one and only structure of a target system and a system has a determinate structure only relative to a certain description." (Frigg, 2006 p. 11)

But there is no reason for which in structural explanation only one of these descriptions can be explanatory. Remember that in the causal view, only one causal history is admitted as

the one describing the "production" of the phenomenon. As a consequence, if a scientific explanation must show the causal history which lead to the production of the explanandum, no plurality of (causal) explanations is admitted. However, this problem does not apply at all to the case of structural explanation. For what structural explanation is concerned, as far as two mathematical models are both good formalizations of the same theory, there is no reason why the existence of different explanations, grounded on different formalism of the same theory, should represent a problem—as far as the considered explanations. To be sure, the availability of two different structural explanations would be a problem for a structuralist, or for a realist on structures, but it is not a problem of structural explanation *per se*, it is instead a problem that concern the general debate realism/antirealism. But as I have already said, following in this Hughes (Hughes, 1993), the problem of realism and explanation should be kept apart, and that a theory on scientific explanation cannot be rejected due to its possible openness to an antirealistic stance.

3.5 Heisenberg's Uncertainty Relations.

As a further argument in favour of this claim, we introduce our second example, that involves Heisenberg's Uncertainty Relations, and that is supposed to show the extent to which structural explanations are *already* used and commonly accepted among physicists.

According to the well-known uncertainty relation between position (x) and momentum (p), the corresponding magnitudes of any quantum system cannot be simultaneously sharp. The product of the uncertainties of the two measurements is formally expressed by the equation:

$$\Delta x \cdot \Delta p \geq \frac{\hbar}{2} \qquad (1)$$

The explanation of the lack of simultaneously possessed values of position and momentum usually invokes the description of a typical "experimental setting" (a thought experiment, really), namely, the measurement of the position of an electron by Heisenberg's microscope. Heisenberg's account of this experiment made use of a qualitative argument, according to which, due to its impact with the gamma ray generated by the microscope "[a]t the instant of time when the position is determined, that is, at the instant when the photon is scattered by the electron, the electron undergoes a discontinuous change in momentum" (Heisenberg, 1927, pp. 174-175).

Given the oversimplified character of this quasi-classical picture of a collision, and the well-known complications of the standard account of measurement, however, it follows that a clear physical explanation of these relations is currently unavailable. Furthermore, "working physicists" often do not regard talks of "discontinuous changes of the wave function" as a realistic description of physical phenomena. Given the current state of affairs, if the availability of a clear physical account were a necessary condition for having some insight into physical phenomena, the Heisenberg's Uncertainty Principle ought to be considered a mystery. However this is not the case: "working physicists" (and not just them) do not regard such a principle as unintelligible.

We argue that this is due to the fact that quantum theory provides us with a solid structural explanation of the uncertainty relations. Within the illustration of the position/momentum uncertainty relation, the physical account depicted above is always accompanied by a much more wholesome formal account. Such an account shows how the formal representative $\Psi(p_x, p_y, p_z)$ of the momentum of the electron is the Fourier transform of the function $\Psi(x, y, z)$, which represents the coordinates giving the position of the particle. Consequently, the structural explanation of Heisenberg's relations exploits the well-known property of the Fourier transform on the basis of which the narrower the interval in which one of the two functions differs significantly from zero, the larger is the interval in which its

Fourier transform differs from zero, in such a way that eq. (1) must be satisfied. The Uncertainty Relation between position and momentum, therefore, is understood as a simple and direct consequence of the mathematical properties of the Fourier transform.

Notice how also this example is a perfect example of the definition of structural explanation provided above. The existence of a minimum for the product of the uncertainties of these two measurements, or the non-simultaneous sharpness possessed by the two observables, is explained by showing how, given the properties of the formal model *M* used by the theory (the Hilbert space of square summable functions) and given the formal representative *A* ($\Delta x \bullet \Delta p \geq \frac{\hbar}{2}$) of Heisenberg's relation (*B*), we have that (i) $\Psi(p_x, p_y, p_z)$ is the Fourier transform of the function $\Psi(x, y, z)$; (ii) for the mathematical properties of Fourier transforms, *A* is satisfied; (iii) the properties of *A* are constrained by the general properties of *M*.

But this is not all. An essential step towards today's understanding of Heisenberg's relations has been taken some years later, in particular with the more general derivation provided in 1929 by Robertson (Robertson 1929). This step was essentially the achievement of a deeper understanding of Heisenberg's principle *via* a different structural explanation. In agreement with Hughes/Clifton's view, we can show that the *new* formal representative of physical systems is

$$\Delta\alpha \cdot \Delta\beta \geq \frac{1}{2}\left|\left(\Psi, [A, B]\Psi\right)\right|$$

(2)

In this formula *A* and *B* are any two non-commuting operators (also spin in different directions, for example), [A, B] is their commutator, so that for every state Ψ, and every pairs of non-commuting observables α and β and corresponding operators *A* and *B*, the product of the uncertainties is greater than the expression on the right hand side. Again, as required, (2) was not an explicit element in the formal models provided by quantum mechanics: the role of

the structural explanation based on Robertson's derivation is exactly that of showing how Heisenberg's Uncertainty Principle is built into the very formal structure of quantum theory, and is therefore part and parcel of these models. It is finally worth noticing how the undeniable generality and insight provided by the Robertson's derivation is independent of any "physical" or intuitive model of the phenomena (as it could have been argued for the case of the interaction photon/electrons in the position/momentum uncertainty relation).

3.6 Structural Explanation and the Other Theories of Explanation

It is worth noticing that the Deductive-Nomological model of explanation fails to account for the examples discussed so far. We have already briefly anticipated this point *a propos* of non-locality: non-local correlations across spacelike separated regions are understood in terms of the properties of their formal representatives in the model − namely the existence of entangled, non-factorizable states. Conservations of physical quantities like angular momentum are not sufficient to explain the correlation of measurement outcome, even if they are part of the explanatory account.

The same point holds for Heisenberg's Principle: in order to derive the position/momentum uncertainty relation we have used the Fourier transform's rule that the narrower is the interval where $\Psi(p_x, p_y, p_z)$ significantly differs from zero, the larger is the interval in which its Fourier transform $\Psi(x, y, z)$ differs from zero, and *vice versa*. Obviously this mathematical "law" corresponds to a physical fact or a physical regularity; however the point here is that in the usual account of Heisenberg relations it is such a physical regularity that is explained/understood in terms of the mathematical law, and not the other way around. It follows that even if we wanted to construct a deductive argument with the formal representative *A* as its conclusion, we could not use it as a prediction, as it is the case with DN

explanations, because in this case reference to initial/boundary conditions would be out of place.

Finally, with respect to the unificationist theory, there is a sense in which the act of "placing the formal representative of the explanandum in the model", or making explicit its role in the model, may help to uncover some relationship to other parts of the model that were previously hidden, thereby diminishing the number of independent elements defining the formal model. But we do not see any necessity in this, and therefore no conceptual link between the unificationist theory and structural explanations. However, whenever deep mathematical theorems succeed in unifying different mathematical theories, any formal representative of a physical explanandum living within one of the unified theories can be "seen under a different light", and can thereby improve our understanding of the physical world by diminishing the number of independent mathematical hypotheses that are indispensable to formulate the theory. In this sense, the unificationist theory of explanation can certainly be accepted within a structuralist explanatory framework, but it seems completely parasitic on the kind of unification realized *within* mathematics.

In the last section we had tried to show how structural explanation is an actual reality in science and how physicists do use the formal resources of scientific theories in order to understand phenomena. So far, however, we haven't faced in details one possible important objection that still can be advanced against the minimal formulation I provided of structural explanation.

At the end of the days, it could be said, the examples of structural explanation proposed in the previous chapter seem to be nothing more than a formal derivation of the explanandum. If this is so, what we are calling structural explanation could end up to be easily accounted for within the good old DN model of explanation. Where, therefore, is the difference between the two kinds of explanations? And why structural should be preferred to DN explanation?

There are different answers which can be opposed to this objection, and each one stresses an important aspect of structural explanation to which we have only alluded in our first characterization. Let's see them in order.

i) first of all it is to be noticed that structural and DN models of explanation meaningfully fit into two different theoretical conceptions of scientific theories. As it is well known, the natural locus for the DN model of explanation is the axiomatic conception of scientific theories, where theories are conceived as a partially interpreted axiomatic calculi—while the natural locus for structural explanation, where models play the central role, is naturally fitted into a semantic view of theories (see Hughes, 1989a). (messo nel capitolo su Hughes) "if one believes (as I do) that scientific theories […] provide explanations, then one's account of explanations will be tied to one's account of scientific theories" (Hughes, 1989a, p.257) In a more general theoretical context where a semantic view of theories is preferred to the received view, the DN account of explanation *at least* looses the privileged role it maintained, which was supported by its naturalness within the received view of theories. In particular, in order to make a covering law model of explanation match with a semantic view of theories the problem is first to be solved of the role of laws of nature within the latter view. The argument just proposed, however, is not a compelling argument for a commitment on the substantial difference between DN and structural explanation. Even if less natural within a semantic view than in the old received view, DN explanation could still be an acceptable account within the semantic view. One way to reconcile the covering law model of scientific explanation and the semantic view of theories could be for instance to consider laws of nature as ruling the behaviour of entities and processes within the model. In this case the DN explanation could still make sense in the same

theoretical context as the one applicable to structural explanation, and therefore could still embed the latter.

ii) We should therefore first of all verify if it is actually true that the examples we proposed of structural explanation can in fact be accounted for within the DN model. Take for instance the example of the Heisenberg Uncertainty relation between position and momentum. Following the DN model of explanation, in order to explain these relations we should be able do deduce them from the initial conditions and some laws of nature. Let's take as initial conditions the description of the state Ψ of the electron with determinate momentum and of the state "ready to measure" of the measurement apparatus. What could we take as the appropriate law of nature in this case? Let's take at face value that von Neumann's projection postulate is a law of nature. This would not help much, for while von Neumann's postulate permits to derivate the appropriate state Ψ' with determined position and undetermined momentum, it still remains to be derived that within each state Ψ' with determined position, the momentum is necessarily undetermined. We hope we have already clearly shown in the previous section how in order to derive this explanandum it is instead necessary to make reference to the *mathematical* properties (laws) of the formal structure whose Ψ' is part—i.e. the properties of Fourier transforms, or, with Dirac notations, the properties of vector spaces. These mathematical laws are therefore essential in order to derive the uncertainty relation, which, again, are not logically derivable from any appropriate law of nature alone. I must insist on this point: it is true that there is a physical regularity that corresponds to the already mentioned property of Fourier transforms, but *that regularity is exactly the explanandum, understood in terms of the properties of Fourier transforms*.

iii) Finally, there is another important aspect where structural explanation differentiates from the DN model. In the previous section we have said that structural explanation provides understanding of a physical explanandum in terms of the formal properties of its representative. This gain of understanding works through a clarification of the role of A within the web of relations that constitute the structure of the theory. That is why it is the very same act of "making explicit" how A is part and parcel of the model (making use of mathematical derivations and proofs) that provides us understanding: it is the use of derivations and proofs exploiting (and therefore putting in evidence) the structural properties of A that clarifies the role of A within the model. Notice that this account of the kind of understanding permitted by structural explanation also explains several significant features of the explanatory power of the mathematical structure of theories, which cannot be explained by the DN model. Under the proposed account, in fact, the evident different explanatory power of different mathematical derivations of the same fact is a natural consequence of the fact that some proofs exploit more than others the underlying structure of the theory; on the other hand, the fact that different formalizations of the same theory can provide us with a significantly different understanding of phenomena depends by the fact that the same element of the theory is described by the two formalism as owning different mathematical properties. This is the case, for instance, of standard quantum mechanics, with the standard vector space formalism, on one hand, and of Quantum Information Theory, with the algebraic approach, on the other. Not only: also the fact that more general derivations (as the one provided by Robertson in 1929) are in general more explanatory is easily accounted for within this picture: the more general is a derivation, the many connections between the explanandum and

> other elements of the theory are individuated, and, in turn, the more deep is the understanding we gain of the explanandum. The DN model of explanation is incapable of making all these distinctions. Within the DN model any logical derivation (any mathematical proof) should be explanatorily equivalent, for if explaining means logically derive, then every logical derivation should be equivalently explicative—instead, within structural explanation logical derivation is one of the tools that can be used in order to highlight the role of the explanandum within the structure of the theory.

This last point give us the opportunity to consider another issue which is intimately linked to the question of the relationship between structural and DN explanations, but which we look anyway as distinct. The problem is that the characterization of structural explanation we have proposed so far seems to be broad enough to take any formal derivation of a physical phenomenon as a structural explanation. This, someone could argue, could represent a problem for structural explanation. Notice that stressing the different explanatory power of different mathematical derivations does not help to settle the problem, for the most part of pure mathematical derivations (which can be more or less "accurate" following the standards we stressed above) are not scientific explanations, period. In other words, while the first is a problem of evaluation of structural explanations, the first concern the recognition of explanations as contrasted to pure descriptions.

In our view this question should be seen as a part of the more general problem of the discrimination between explanations and simple descriptions—the *illocutionary force problem*, as called by Achinstein, is therefore not peculiar of structural explanation, but concerns many theories of explanation (see Achinstein, §3).

With respect to this point, we think that the fundamental required discriminating factor between the two objects of science is to be found within pragmatic, that is within the scope which explanations and descriptions serve to.

More in particular, it can be argued with Achinstein that a suitable characterization of the concept of explanation cannot be independent from the illocutionary act which produces it, that is the act of explaining, where for our purposes we can be content with Achinstein's minimal condition for an explanatory act: "S explains q by uttering u only if S utters u with the intention that his utterance of u render q understandable" (Achinstein, p. 16) Achinstein's view of explanation holds that one "must begin with the concept of an illocutionary act of explaining and characterize explanations by reference to this, rather than conversely" (Achinstein, p.102).[12] A necessary condition for the explanatory nature of u, therefore, is not given by any intrinsic feature of u, but by its pragmatic role of standing in the relation "being an explanation of" with a given explanandum q.[13]

Back to structural explanation, we think this general account of the illocutionary force problem should be enough to settle also the problem of the distinction between those cases in the history of science where a mathematical derivation can be considered a structural explanation and those where not.

[12] Holding this view, however, implies neither that for every single utterance of u it corresponds an explanation, nor any ontological stance about explanations as sentences or propositions. For an account of how avoiding these and other possible problems, see Achinstein (1983, § 2, 3).

[13] For the other sufficient conditions that u must satisfy in order to be an explanations, we also refer to Achinstein (1983, § 2).

Chapter 4

Models

4.1 Introduction

We have seen so far that structural explanation exploits the resources of mathematical models displayed by highly abstract theories. Obviously given that we have called models into question, we need some account of what models are and what kind of knowledge they permit when they are used for an explanation. The problems concerning scientific models are well expressed by Hughes: "[o]ne major philosophical insight recovered by the semantic view of theories is that the statements of physical theory are not, strictly speaking, statements about the physical world. They are statements about theoretical constructs [i.e. models]. If the theory is satisfactory, then these constructs stand in a particular relation to the world. To flesh out these claims, we need to say more about what kinds of constructs are involved, and what relation is postulated between them and the physical world". (Hughes, 1997, p.325)

This chapter is dedicated to the attempt to provide such a theoretical background to structural explanation.

As a start, if models are meant to teach something about a target, they must represent—however, as far as we don't have an account of how and in virtue of what scientific models represent, the above reformulation of the problem does not help much. On the other hand, we do have now a useful reformulation of the question about what models are, and what kind of knowledge of the target they support into the question: what is scientific representation, and in virtue of what a model is said to represent its target? In the following it will be evident that I have tried to put forward an analysis of scientific representation which does not compel to any realistic claim about science. However, such an attitude is not to be taken as an antirealistic stance towards models or science in general, since I think that, as in the case of structural explanation, it is a virtue of a coherent account of scientific representation that it keeps apart the debate realism/antirealism. On the other hand, not only I maintain that all that will be said in the following is coherent with realism, but that the proposed neutral characterization of scientific representation and of the relation between the latter and explanation could provide a useful point of departure for a formulation of a realist stance.

A caveat, before going on: in the following discussion I am not meant to achieve a full fledged theory of models, or of representation. My much more modest aim is to concentrate on some characteristic on models, that are rightly put in evidence by the theories that I will discuss, and that will shed some light into the proposed theory of structural explanation. Hopefully, this discussion will lead to new cues for a full understanding of the latter.

4.2 Outline of a strategy

Roman Frigg (2006) individuates three conundrums that a theory on scientific representation must answer, one of ontological and two of semantic nature. The first problem, which he calls the '*ontological puzzle'* concerns what kind of entities are scientific models:

"Are they structures in the sense of set theory, fictional entities, concrete objects, descriptions, equations or yet something else?" (Frigg, 2006, p. 2). The first semantic problem is referred by Frigg as the '*enigma of representation*' and consist in the question how do scientific models (whatever they are) represent their target. Finally, the third problem is the '*problem of style*' which comes in a factual and a normative variant: the first consists in the fact that a theory of representation must be able to embed all different styles of representation present in science (mathematical, analogical, scale models, idealized..); the second consists in the fact that arguably not all the possible styles of representation can be considered scientific representation—our discussion must therefore also address this problem and the question of which are, if there are, the norms which discern scientific from other representations. Notice that, although Frigg keeps care not to conflate them, the issues of the enigma of representation, and that to the normative version of the problem of style can be easily overlapped. If the enigma of representation consists in how *scientific* (not generic) models come to represent, it is possible that the answer has something to do with the specific style that scientific representation takes. Conversely, if the style of a scientific representation does not bear any relation to the answer to the enigma of representation, then arguably the is not specific of *scientific* models. I'm not arguing here for a necessary conflation between the enigma of representation and the puzzle of style. However, taking into account the possibility of such a conflation will be useful later, when we will try to analyze how the illustrated theories on representation answer to these problems.

Nothing will be said in the following about the ontological puzzle—my main concerns in this chapter will be the semantic problems, given their relevance to the relation between models and explanation.

Before concluding with this introduction to the problem I must list two requirement that must be satisfied, following Frigg, by an acceptable theory on scientific explanation: the first

is it must reflect that scientific models represent their targets in a way that allows us to acquire knowledge about them. This requirement corresponds to what we will call, following Chris Swoyer, *surrogative reasoning* (Swoyer, 1991). The second requirement is that our theory must allow misrepresentation, i.e., explain how is it that scientific models (as, for instance idealized models) often misinterpret parts of the target, and still are good representations.

In the following we will present three theories on scientific representation that pivot on two key concepts. First of all if *a* represents *b*, then *a* stands for, refer to *b*. Secondly, the notion already introduced of surrogative reasoning: scientific representation must allow an informed user to draw inferences about the represented object.

The first two theories, R.I.G. Hughes' Denotation, Deduction, Interpretation theory (DDI) (Hughes, 1997) and Mauricio Suárez's inferential theory (Suárez, 2004) may be said to escape to provide a full answer to the three conundrum, for different reasons. The first because it is not meant to be a complete theory on scientific representation, but the settlement of some basic concepts that should be present in a full fledged theory. Suárez's theory, instead, expressly rejects the possibility of providing a substantial account of scientific representation, claiming that the latter is a notion that does not allows necessary and sufficient conditions. As such, a theory can at best individuate a set of necessary but not sufficient conditions for scientific representation (so no full answer can in principle be given to the enigma of representation) and, as a consequence, also the normative version of the problem of style can have an only partial answer, given by a set of necessary conditions. The third theory presented is Gabriele Contessa's interpretational theory (Contessa, 2007). This theory is explicitly inspired by Suárez theory in the role that the latter entrusts to surrogative reasoning, but it intends to provide a substantial account of representation, which can explain what is the link between representation and surrogative reasoning.

After an illustration of Hughes' (as usual) pivotal work, I will analyze the main points and the answers to the semantic questions displayed by Suárez's and Contessa's theory. In § 3 I

will argue that the inferential theory has many undeniable merit: it points out that truthfulness is not a necessary feature for scientific representation, it recognizes the importance of contextual factors in the determination of the conditions for scientific representation, and it recognizes that scientific representations have an essentially non arbitrary feature. However, on the one hand I show that Suárez fails to capture the very nature of scientific representation's objectivity (due to the attempt to provide a non contextual characterization of such a notion) on the other hand I show that the inferential theory does not help much to solve the conundrum of scientific representation, given that, rather than explain the principal features of scientific explanation, it merely states them. For what the interpretational theory is concerned (§ 4), it surely tries to surpass inferentialism by providing a more fundamental theory which could explain the relation between surrogative reasoning and scientific explanation. However, I argue that, due to the defection of the central concept of objectivity and of the contextuality of representation, Contessa's theory leads to paradoxical implications. Finally, I also argue against Contessa's claim that interpretation is more fundamental than surrogative reasoning, and that the latter occurs in virtue of the former. The analysis of Contessa's and Suárez's theories and of their shortcomings will lead us to another proposal (§ 5), consisting first of all in a new contextual formulation of the concept of objectivity according to which a scientific model is an objective representation relative to a determinate context of inquiry C iff it support the achievement of the epistemic aims of C.

4.3 Denotation, Demonstration and Interpretation.

I alluded at the end of the previous section that our discussions will mostly concern Suárez's and Contessa's theories, given the reflection about their merits and shortcomings will finally

lead to the proposal I will put forward. Still, Hughes' theory deserves one section first of all for the clear influence that this theory have on Suárez's and Contessa's—moreover because the analysis of the DDI theory will someway complete our review of Hughes' conception of structural explanation. According to the DDI account (Hughes, 1997) "[t]he characteristic—perhaps the only characteristic—that all theoretical models have in common is that they provide representations of parts of the world, or of the world as we describe it." (p.325). However, on the one hand not all representations are scientific models (think, for instance to Vermeer's canvas "View of Delft"), on the other hand, the kind of representation provided by scientific models is in no obvious sense every time the same. So Hughes' question is: "[w]hat […] does the representation of a crystal as an arrangement of rods and spheres have in common with the representation of the motion of a falling body by the equation $s=gt^2/2$?" (pp. 325-326).

The use of a model in science involves three components: Denotation, Demonstration and Interpretation. Denotation defines the relation between the model and the target system in such a way that the model is a "symbol for it, stand for it, refer to it" (p.330, quoted from Goodman's 1968). With respect to denotation, Hughes recognizes two genres of models: *local* and *global*. Local theories (and the local models they identify) deal "with a clearly specified type of physical system" (e.g. Bohr's theory of hydrogen atom, or the Cosmo), while global theories identify a whole class of models and is essentially "a set of instructions for building a great variety of local models" (e.g. Newtonian mechanics). Notice here that in stating the rule "no representation without denotation" Hughes rules out as a fact the possibility of representation of non existing objects. This will be one of the reasons which Suárez (2004) adduces for the need to overcome the DDI theory.

The second stage of Hughes' account of models is Demonstration. With an illustrating example Hughes claims that "models are always *representation as*[14], representations of the kind exemplified by Joshua Reynolds' painting of Mrs. Siddons as the Muse of Tragedy. In this painting Reynolds invites us to think of his primary subject, Mrs. Siddons, in terms of another, the Muse of Tragedy, *and to allow the connotations of this secondary subject to guide our perception of Mrs. Siddons*" (p.331) (mine italics). The last part of this quotation refers to the fact that "from the behaviour of the model we can draw hypothetical conclusions about the world over and above the data we started with."(p.331). This is the stage of Demonstration, that takes care of what we call surrogative reasoning. In order to support Demonstration, therefore, both analogical and abstract mathematical models must have *an internal dynamics which is independent on the known structure of the studied system* but which permits to the subject to perform deductions and therefore draw conclusions about the system. As an example, take the two-slit interference. We can model it either with the mathematical model of wave functions, or with a real ripple tank. Both models have an internal dynamics, represented in the first case by geometry and algebra, that of the ripple tank by the physical processes which are involved in the propagation of water waves. Both dynamics permit us to conclude that the "distance between interference fringes varies inversely with the separation of the sources, and also with the frequency of the waves". Notice that the importance of the requirement that the internal dynamics of the model must be *independent on the known structure of the studied system* is given by the fact that such a dynamics must permit the inference not of merely new, but novel predictions. If the model's dynamic is completely parasitic on what is known of the represented system, then this will not

[14] As his theory of explanation, so Hughes' theory of modelling is, we think, inspired by Nancy Cartwright's work. In Cartwright's theory the process of relating theory with phenomena comes in two stages (see Cartwright, 1983, p.133-134). We begin by an unprepared description, let's say a collection of all the gathered data. The first stage of theory entry is 'preparing the description'. Here the phenomenon is presented in a way that will 'bring it to the theory' (p.133) and the subject begins to refer to the physical system in terms of the model (i.e. she talk of the system 'as if' it is the model). The second stage of modelling is the stage in which laws are applied to the model.

able to support novel predictions. We add here that the possess of such an internal dynamics is necessary to allow not only novel predictions, but any kind of surrogative reasoning. In particular, for someone, like me and Hughes[15], endorsing a conception of explanation where the explanatory burden is held by models, the model must provide the means for the explanation to be possible. If, for instance, the model is supposed to be utilized in a causal explanation of a given phenomenon, the former must exhibit some causal relations between his elements.

In Suárez (2004) it is argued that the DDI account is undermined by the fact that according to it representation requires the actual carrying out of inferences about the target on the part of an agent. This is one of the points that, following Suárez, distinguish the DDI from his own theory. But why should the fact that representation requires demonstration essentially, be an issue in Suárez's view? On the one hand, if Suárez's worry is that the requirement of demonstration be flattened to the mere act of a user, then this is obviously not the case in Hughes' theory. On the contrary, throughout Hughes' discussion, Demonstration has been put forward as a requirement concerning an *internal structure* of scientific models, which contains "resources which enable us to demonstrate the results we are interested in" but which is independent by any actual inferences carried out by a user. This is also testified by the above reported requirement on the dynamics of the model, i.e. the requirement that the model to be independent on the known structure of the studied system. On the other hand, if Suárez's intention is to guarantee that the concept of representation be captured by some relation between the target and the vehicle alone (so, in particular, to guarantee that the representation is a concept independent by any agent or action), then this worry is in open contrast with Suárez's own characterization of representation. In order to understand why, we need to anticipate something about Suárez's theory. A central notion in Suárez's account of representation is the notion of *representational force*, or simply *force.* This is "the capacity of

[15] But also Woodward (2003), van Fraassen (1980) and Cartwright (1983).

a source to lead a competent and informed user to a consideration of the target. Force is a relational and contextual property of the source, fixed and maintained in part by the intended representational uses of the source on the part of agents: No object or system may be said to possess representational force in the absence of any such uses." (Suárez, 2004, p. 4). It is a necessary condition for representation that the representational force of the target points to the vehicle. But if it is legitimate to talk of force only in the presence of a determinate context of use, then also Suárez's account of representation depends on the context of use (and not uniquely on a relation between target and vehicle).

A final way of reading Suárez's criticism could be this: representation is a contextual notion, but it is not dependent on the *actual performance* of a surrogative reasoning on the part of a user. This is in fact a legitimate requirement, but, again, I doubt that it is coherent with Suárez's own theory. The reason for this is that, although force is undeniably a fuzzy concept, its characterization seems to imply essentially the practical use of the model by part of one or more agents.[16] [17] It is dubious then that an eventual appeal to the fact that such uses are not the surrogative reasoning which was the subject of Suárez objection could serve in any way to solve the contradiction.

Let's now come back to the DDI theory. We have said that the deductions performed within the model permit us to draw conclusions about the subject. This, however, is not completely true at the stage of Demonstration. The results obtained within the model at this stage are, instead, only results about the model itself (wave functions or ripples). In order to be able to draw conclusions about the system, the subject must therefore interpret the obtained results in terms of the system. Within the stage of Interpretation the agent sees it to that the results gained within demonstration are finally taken, as Hughes says, 'back to the world'. Moreover,

[16] This seems presupposed in the previous quotation but also when Suárez says that the requirement of force "leaves open the issue of how many agents are required in a scientific community to fix the representational force of a source" (ibid. p. 10)

[17] For that matter, not only I see no problems in acknowledging that representation implies demonstration essentially, but this will be the main point I will argue for in my criticism to Contessa's theory.

it is at this stage that the empirical adequacy of the model is verified, by checking whether the inferences carried at the stage of demonstration are true in the world.

4.4 The inferential conception.

The DDI account just illustrated was not, in Hughes' intention, to be conceived as a general theory of representation—nor the three components of Denotation, Demonstration and Interpretation were conceived by Hughes' as individually necessary or jointly sufficient conditions for representation. The more modest suggestion advanced was instead that "if we examine a theoretical model with these three activities in mind, we shall achieve some insight into the kind of representation that it provides" (p.329).

Arguably, in the eyes of Mauricio Suárez, Hughes' first error is this, i.e. the idea that a theory on scientific representation could individuate general necessary and sufficient conditions for it: "[r]epresentation is not the kind of notion that requires a theory to elucidate it: there are no necessary and sufficient conditions for it"—at most, a general theory of representation can provide necessary conditions for it.

In Contessa (2007) it is argued that it is not clear what reasons led Suárez to such a position. Following Contessa, it seems sometimes that the possibility of general, necessary and sufficient conditions is rejected due to a conception of representation as a non general notion—which has just general necessary conditions and that meets different stronger conditions dictated contextually by the use it is made of the model. In this case the inferential theory would be in fact a substantial but minimalist theory, which could be further developed in order to find what sufficient conditions are met in different context of inquiry. I think that this reading of Suárez's position can be dismissed. Suárez never says that the further contextual conditions that will be met will be *sufficient* conditions. Suárez's motivation for

rejecting the research of a “general, necessary and sufficient conditions” seems, clearly enough, to be that there are in fact "no deeper features to scientific representation other than its surface features"—i.e. denotation and surrogative reasoning. The inferential theory is therefore both a deflationary and a minimalist account of representation.

Another point of departure of Suárez’s discussion is the distinction between a representation and an accurate or truth representation: “the primary question is not: “how does the graph manage to represent the bridge accurately or truthfully? But rather: “in virtue of what is the graph [of the Forth Rail bridge] a *representation* (however accurate or inaccurate) of the bridge?””

4.4.1 Objectivity.

Representation is therefore characterized within the inferential account by the following criterion:

[inf]: *A represents B only if (i) the representational force of A points towards B, and (ii) A allows competent and informed agents to draw specific inferences regarding B.*

(i) can be paraphrased as the condition that the model is used by someone to represent the system (Contessa, 2007). The *representational force* of a source (or, more simply, its *force*) is “the capacity of a source to lead a competent and informed user to a consideration of the target”. Remember its characterization as “a relational and contextual property of the source, fixed and maintained in part by the intended representational uses of the source on the part of agents: No object or system may be said to possess representational force in the absence of any such uses”.(p.4)

As already recognized by Hughes (and by Hertz before him), Suárez considers the notion of surrogative reasoning as central in scientific representation. Scientific models do more than merely denote an object. Differently from generic stipulation, scientific representation permits to draw relevant conclusions about their target: they are *informative*

about it. It is in this sense, that is in the sense that they are informative, that Suárez's claims that scientific models *objective*.

> "[inf]'s part (ii) has the important function of contributing the *objectivity* that characterises scientific representation. In contrast to part (i), it in no way depends on an agent's existence or activity. It requires A to have the internal structure that allows informed agents to correctly draw inferences about B, but it does not require that there be any agents who actually do so. And this turns out to be exactly the feature that distinguishes cases of objective scientific representation (however inaccurate) from ordinary representation by arbitrary stipulation." (p.12)

If a theory on scientific representation is meant to account for the deep grounded intuitions that are linked to the notion of representation, it must also account for the fact that scientific representation *is not the product of an arbitrary convention between agents*, and that there must be something in a vehicle V such that a user can legitimately take V as representing a determinate target. I think this is the core of the notion of objectivity that plays a crucial role in the inferential theory.

Consider a peace of paper and two pens writing on it, and stipulate that they represent respectively the sea and two ships sailing on it. Compare then this kind of representation with the opposite in which the paper represents the ships and the pens represent the sea. Suárez claims that "the ships-on-sea system is more "objectively" characterised by the first denotational arrangement than by the second" and with this he means that the second representation "is certainly less informative, since the relative movements of pens and paper can not allow us, for instance, to infer the possibility that the two ships may crash" (p. 8).

Now the problem is how to capture the notion of objectivity. Suárez quickly consider the possibility that dyadic relations such as correspondence truth, isomorphism or similarity could serve this purpose, but he rejects it because according to him they are unapt to ground a

theory of scientific representation in the first place. This is due to the fact that the above mentioned relations fails to satisfy some formal requirement that a characterization of representation must meet.[18]

Suárez proposal is therefore that it is exactly the capacity to allow surrogative reasoning that determines the objectivity of scientific models

To see how Suárez proposal cannot be satisfactory, consider again the example of the ships on the sea. If it is true that with the representation: ships=paper, pens=sea, you cannot draw the conclusion that the two ships may crash, you can draw as many other conclusions than with the other representation. You can for instance conclude (under the assumption that the paper is rigid) that whatever movements the ships can perform, they will always remain at the same distance one from the other, or that if you burn one of them also the other will burn. These conclusions are obviously false, but if truthfulness and accuracy are irrelevant for objectivity, then this fact cannot be used in order to argue that one representation is more objective than the other.

Take the other example that Suárez uses as an illustration of objective difference between simple representation by stipulation and scientific representation: the representation of the city of Edinburgh by the Forth Rail Bridge:

> "Consideration of a graph of the Forth Rail Bridge does not lead an agent to any reliable conclusions regarding the city of Edinburgh (i.e. conclusions that could not have been derived on the basis of a consideration of any other object)."

So, from an analysis of the bridge, we can calculate that, say, the outside double-cantilever shoreward ends carry weights of about 1000 tons to counter-balance half the weight of the suspended span and live load (see http://www.forthbridges.org.uk/railbridgemain.htm). We can therefore draw the conclusion that the shoreward part of the city of Edinburgh carries

[18] For a discussion of these requirements see (suarez ???) but also Frigg (2006???)

1000 tons to counter-balance half the weight of downtown. *The conclusion is obviously not true, however, it is reliable in Suárez's sense that it could not have been derived on the basis of any other object* (which, true, is anyway a curious definition of reliability).

If this is so, how are we supposed to discriminate between objective and non objective representation? The above examples would suggest that it is in fact the truthfulness of the surrogative reasoning that measures the objectivity of a model—but Suárez rejects by the very begin this alternative. He is right, anyway, to do so, given that if objectivity would be defined in terms of truthfulness or accuracy, it would be very difficult (if not impossible) to account for misinterpretation. On the other hand the notion of surrogative reasoning must be obviously further characterized, because the requirement that a user can draw inferences about the target is obviously too weak to grasp the objectivity of representation. Summing up, Suárez's failure in defining objectivity is due to the friction between two factors: on the one hand Suárez guesses that surrogative reasoning is at the core of objectivity, on the other hand, he cannot find any criterion which further classifies surrogative reasoning, in such a way to rule out as non-objective cases like the representation of the city of Edinburgh with the Forth Rail bridge. I will argue in §??? that the required narrowing criterion for surrogative reasoning is to be found in the precise kind of use that it is to be made of the model. The objectivity of a representation, therefore, is not captured by a dyadic relation like isomorphism, similarity and so on, but by a triadic relation between the target, the vehicle and the context of inquiry: the vehicle must support (say) a surrogative reasoning amounting to a causal explanation of some part of the target.

4.4.2 Contextuality

Now, we have said that the inferential theory cannot provide a full answer to Frigg's questions, however, it will be useful to spell out with more clarity the way Suárez's theory

deal with the enigma of representation and the problem of style. Although in fact the deflationist character of the inferential theory implies that there are no necessary and sufficient conditions for representation, the theory does provide some necessary conditions: [inf]'s (i) and (ii). Arguably both (i) and (ii) are principles which are meant to define a partial answer to the enigma of representation. That is, they do not merely individuate the conditions for a scientific representation to be acceptable, but they are the very necessary conditions for a vehicle to represent a target. It could be argued that while (i) is clearly a condition for representation, (ii) seems to be more a criterion for the style of a scientific representation. Since Suárez does not directly address Frigg's questions it is not easy to say what he precisely has in mind. On the other hand, I have already noticed in the introduction that the enigma of representation and the problem of style can possibly overlap, and the same requirement can serve to answer to both the questions. In other words, [inf]'s (ii) could be both a necessary condition for a scientific model to represent, and a condition on the style of scientific representation.

Now, besides the fact that Suárez's [inf] is not concerned to provide sufficient conditions for representation, however, I think that [inf] can achieve much less than the same Suárez was intended to. The problem, again, is in the concept of objectivity. The declared intent of [inf]'s (ii) is that of ruling out those kinds of interpretations which are not objective, not informative. However, I have argued that, as characterized by Suárez, surrogative reasoning, however, cannot discriminate with those cases which are intuitively objective and those that are not. As a conclusion, [inf] does not manage to guarantee for a feature of scientific representation that Suárez consider essential.

At the end of the previous section I have argued that in order to capture the notion of objectivity it is necessary to take into account also the context of inquiry in which this takes place. On the other hand, Suárez did recognize the important role of the context of inquiry in the realization of scientific representation. We have already seen how the representational

force of a source is "fixed and maintained in part by the intended representational uses of the source on the part of agents: No object or system may be said to possess representational force in the absence of any such uses." More exactly, following Suárez it is the context of inquiry that determines the additional necessary conditions, other than [inf]'s (i) and (ii), that a representation must meet. These conditions, says Suárez, can be isomorphism, similarity, or others. Now, I argue that Suárez fails to capitalize on this intuition about the importance of the context in scientific representation. The problem is that while he admits that the context of inquiry decides the additional necessary conditions for representation, he does not say why or how. He never specifies in which way should a determinate context of inquiry impose additional conditions to the basic one. The obvious answer, as I argue, is that to any specific context of inquiry a determinate surrogative reasoning, with determinate features will be required. A formal (non interpreted) model, for instance, would not be able to support the causal reasoning constituting a causal explanation. In this sense, the context of inquiry of a causal explanation add the constraint that the model must have an internal causal dynamics. If these conditions are not satisfied, the model is not adequate in that activity, and the model therefore is not an objective representation.

As a final remark, notice that there is another reason for which the inferential theory could be considered not fully satisfactory. Besides the fact that its answer to the enigma of representation is unsatisfactory, the problem is that Suárez's theory does not *account* for the factual aspect of the problem of style, for the relation between scientific representation and surrogative reasoning and for misrepresentation: *it merely states them*. Probably this is a consequence of the deflationary character of the theory, and of the claim that representation does not have other features but its surface features. However, undeniably this lead to a quite limited understanding of scientific representation.

4.5 The Interpretational Conception.

Gabriele Contessa's interpretational conception of scientific representation (Contessa 2007) starts with the acknowledgment of two important merits of the inferential conception: the distinction between representation *simpliciter* and accurate, true and complete representation—and the elucidation of the importance of the relation between scientific representation and surrogative reasoning. However, Contessa rejects the deflationary character of the inferential theory. Against Suárez's conception Contessa proposes the interpretational conception as a *substantive conception*, a characterization of representation providing both necessary and sufficient conditions for representation. In Contessa's theory therefore, we should finally find an answer to Frigg's enigma of representation.

The starting point of Contessa's argument is surrogative reasoning. The very first characterization of scientific representation he provides is that *surrogative reasoning is a sufficient and necessary condition of representation*. This starting point is very important in all the subsequent arguments, for they are expressly conceived in order to explain what Suárez didn't explained: what is the relation between scientific representation and surrogative reasoning.

4.5.1 Interpretation

At the foundation of Contessa's theory is the concept of *interpretation*.

Following a general definition 'a user interprets a vehicle in terms of a target if she takes facts about the vehicle to stand for (putative) facts about the target'. Contessa defines a particular kind of interpretation, which he considers arguably the principal kind of interpretation used in science, although not the only one.

Following Contessa's formulation:

An analytic interpretation of a vehicle in terms of the target identifies a (nonempty) set of relevant objects in the vehicle $\left(\Omega^V = \left\{o_1^V, ...o_n^V\right\}\right)$ and a (nonempty) set of relevant objects in the target $\left(\Omega^T = \left\{o_1^T, ...o_n^T\right\}\right)$, a (possibly empty) set of relevant properties of and relations among objects in the vehicle $\left(P^V = \left\{{}^nR_1^V, ..., {}^nR_m^V\right\}\right)$, where ${}^nR$ denotes an *n*-ary relation and properties are construed as 1-ary relations) and a set of relevant properties and relations among objects in the target $\left(P^T = \left\{{}^nR_1^T, ..., {}^nR_m^T\right\}\right)$, and a set of relevant functions from $(\mathbf{\Omega}^V)^n$—that is, the Cartesian product of $\mathbf{\Omega}^V$ by itself *n* times—to $\mathbf{\Omega}^V$ } $\left(\Phi^V = \left\{{}^nF_1^V, ..., {}^nF_m^V\right\}\right)$, where ${}^nF$ denotes an *n*-ary function) and a set of relevant functions from $(\mathbf{\Omega}^T)^n$ to $\mathbf{\Omega}^T$ $\left(\Phi^T = \left\{{}^nF_1^T, ..., {}^nF_m^T\right\}\right)$

A user adopts an *analytic interpretation* of a vehicle in terms of a target if and only if:

1. The user takes the vehicle to denote the target,
2. The user takes every object in $\mathbf{\Omega}^V$ to denote one and only one object in $\mathbf{\Omega}^T$ and every object in $\mathbf{\Omega}^T$ to be denoted by one and only one object in $\mathbf{\Omega}^V$,
3. The user takes every *n*-ary relation in P^V to denote one and only one relevant *n*-ary in P^T and every *n*-ary relation in P^T to be denoted by one and only one *n*-ary relation in P^V,
4. They take every *n*-ary function in $\mathbf{\Phi}^V$ to denote one and only one *n*-ary function in $\mathbf{\Phi}^T$ and every *n*-ary function in $\mathbf{\Phi}^T$ to be denoted by one and only one *n*-ary function in $\mathbf{\Phi}^V$.

Finally, here's the characterization of representation provided by the interpretational theory:

> 'a certain vehicle is an epistemic representation of a certain target (for a certain user) if and only if the user adopts an interpretation of the vehicle in terms of the target.

An analytic interpretation underlies the following set of inference rules:

Rule 1: If o_i^V denotes o_i^T according to the interpretation adopted by the user, it is valid for the user to infer that o_i^T is in the target if and only if o_i^V is in the vehicle,

Rule 2: If o_1^V denotes o_1^T , . . ., o_n^V denotes o_n^T, and ${}^nR_k^V$ denotes ${}^nR_k^T$ according to the interpretation adopted by the user, it is valid for the user to infer that the relation ${}^nR_k^T$ holds among o_1^T, . . ., o_n^T if and only if ${}^nR_k^V$ holds among o_1^V, . . ., o_n^V,

Rule 3: If, according to the interpretation adopted by the user, o_i^V denotes o_i^T, o_1^V denotes o_1^T, . . ., o_n^V denotes o_n^T , and ${}^nF_k^V$ denotes ${}^nF_k^T$, it is valid for the user to infer that the value of the function ${}^nF_k^T$ for the arguments o_1^T, . . ., o_n^T is o_i^T if and only if the value of the function ${}^nF_k^V$ is o_i^V for the arguments o_1^V , . . ., o_n^V .'

4.5.2 Surrogative reasoning

According to Contessa, within Suárez's theory the possibility of performing inferences from a vehicle to the target 'seems to be a brute fact, which has no deeper explanation'. 'This makes the connection between epistemic representation and valid surrogative reasoning needlessly obscure and the performance of valid surrogative inferences an activity as mysterious and unfathomable as soothsaying or divination' (p.9). A crucial declared merit of the interpretational conception of representation, therefore, is that *it explains why a scientific model permits surrogative reasoning*. In order to see how this explanation works we must first introduce few other notions and some terminology.

With an example taken by Contessa, we can say that both the map and the logo of the London underground represent in a sense the London underground. Following Contessa's terminology, both the logo and the map *denote* the London underground—in this sense of

representation, for an object A to represent an object B it is sufficient that a group of users implicitly or explicitly agree that A does so.

However, the map does not only denote the London underground, but it permits an informed user to perform surrogative reasoning and therefore draw conclusions (although not necessarily sound) about the underground network.

This is possible in virtue of the fact that the map of the London Underground is an *epistemic representation* of the London Underground, where

> "A vehicle is an *epistemic representation* of a certain target for a certain user if and only if the user is able to perform valid (though not necessarily sound) surrogative inferences from the vehicle to the target. I will call this necessary and sufficient condition for epistemic representation *valid surrogative reasoning*."

Notice that following the above characterization, "(e)pistemic representation is not a dyadic relation between a vehicle and a target but a triadic relation between a vehicle, a target, and a (set of) user(s)." (p.4) However, as within the inferential view, in order for a vehicle to be an epistemic representation, there is no need that a user would ever *actually* perform some surrogative reasoning about the target, provided that she would be able to do it in case she is asked to: within the interpretational view surrogative reasoning is nothing more than a symptom of scientific representation.

In order to define what a *valid* surrogative reasoning is, we must come back now to the concept of interpretation.

> 'If a user adopts an analytic interpretation of the vehicle, then an inference from the vehicle to the target is *valid* (for that user according to that interpretation) if and only if it is in accordance with Rule 1, Rule 2, or Rule 3' (p.10)

Finally, analogously to Suárez's distinction between representation and "accurate, true and complete representation", Contessa distinguishes between *sound* and *unsound* surrogative reasoning. A surrogative reasoning is *sound* if it leads to consequences which are true of the target, and then uses the concept of surrogative reasoning in order to classify two different

ways of representation. A vehicle, therefore, can be a *completely faithful*, a *partially faithful* or a *completely unfaithful epistemic representation* of a target if and only if the vehicle is an epistemic representation of the target and, respectively, all, some, or none of the valid inferences from the vehicle to the target are sound. However, for an object to be an epistemic representation of the target, the only requirement is that it permits the performance of *valid* (and not necessarily sound) surrogative reasoning.

We are now able to summarize how and if Contessa's theory answers to the semantic conundrums of representation listed by Frigg. The interpretational theory's answer to the enigma of representation is straightforward: a necessary and sufficient condition for scientific representation is the interpretation of the target in terms of the vehicle. Interpretations in science are typically (but not necessarily) analytic representations. It is also within the notion of interpretation therefore that we find an answer to the normative version of the puzzle of style: what distinguishes scientific representation from the generic representations is the fact that the former are epistemic representations, that is they are interpreted in such a way that they support a set of rules for valid surrogative reasoning. In the following I will challenge Contessa's theory and I will articulate my argument in three points. First, Contessa's demonstration that interpretation is sufficient for interpretation is shown to be inconclusive. Second, it will be argued that the interpretational theory is too inclusive, since from the claim that interpretation is necessary and sufficient condition for scientific representation, it follows that anything can scientifically represent something else—conclusion which is in evident contradiction with the concept of scientific representation that we want to capture. Third, Contessa's claim that interpretation founds surrogative reasoning, and it is therefore this notion that defines scientific representation, is shown to be unwarranted.

4.5.3 Substantialism.

In this section I will first analyse Contessa's argument for substantialism, and prove that it is ineffective. Then, in the next section I'll show that the interpretational theory is too inclusive and leads to paradoxical conclusion with respect to what to count as a scientific representation.

In order to demonstrate that interpretation is sufficient for scientific representation, Contessa proposes to show how its negation is untenable. The best way to argue that interpretation is not sufficient for scientific representation, Contessa argues, is to produce an example in which this is the case. Let's therefore interpret a hockey puck sliding on the surface of a frozen pond in terms of Rutherford's model of the atom. Let's then say that the puck and the surface of the ice are represented within the model respectively by the electron and the nucleus. "[A]ccording to the general interpretation, it would then be possible to infer from the model that, say, the puck is negatively charged and the ice is positively charged, that the puck orbits around the ice surface in circular orbits, etc." (p.11).

It can be argued that the above model is not a scientific representation of the puck on the ice, and that, therefore, the example shows how interpretation is in fact not sufficient for interpretation. According to Contessa, however, this objection is based on a conflation of the notions of epistemic representation and partially faithful epistemic representation. In other words, Contessa assumes that this criticism is grounded on the fact that from the Rutherford model no sound inferences can be derived about the puck on the frozen pond.

On the other hand, if it is so the above criticism does not affect the inferential conception, since within it there is no requirement for the soundness of the surrogative reasoning supported by an epistemic representation. This is how the argument goes: someone denying that interpretation is sufficient for representation must *either* deny that interpretation is sufficient for surrogative reasoning, *or* deny that surrogative reasoning is sufficient for epistemic representation. Given his previous detailed analysis of the relation between

interpretation and surrogative reasoning, Contessa argues that the first hypothesis can be reasonably dismissed and that the only alternative to discuss is therefore the second. The argument is then concluded by considering three possible ways to argue for the fact that, within the example of the puck-on-ice, surrogative reasoning is not sufficient for representation: the first is that surrogative reasoning is insufficient because *all* the conclusions about the system from the model are false; the second is that the Rutherford model is not an epistemic representation of the puck-on-ice system not because all the inferences from the model to the system are unsound but because the user knows them all to be unsound; the third is that the Rutherford model is not an epistemic representation of the puck-on-ice system because no actual user of the model can truly believe that the model allows any sound inference about the target system.

All the three arguments are showed to be untenable, and with this Contessa claims to have demonstrated substantialism. Given the discussion provided before of the contextual character of representation within inferentialism, it is easy to see why Contessa's argument is a little tricky. He takes for granted that the only way that a deflationist could possibly go in order to refuse the puck-on-ice example as a scientific representation must be based on a conflation between notions of epistemic representation and partially faithful epistemic representation. An advocate of Suárez's deflationist view, on the other hand, could consider Contessa's example as meaningless in the first place. Remember that according to Suárez the additional necessary conditions that *scientific* representation meets are contextual, determined case by case by the context of inquiry. If this is so, the non sufficiency of surrogative reasoning (or of interpretation) for scientific representation cannot be proved unless the model is put in a determinate context of inquiry.

4.5.4 Objectivity

The way Contessa dealt with the example of the puck on the ice shows how within the interpretational theory, anything can be taken as a scientific representation of anything else, provided that an interpretation of the latter in terms of the former takes place. Scientific representation, in other words, is characterized as a totally arbitrary notion, which merely requires intentionality and a set of rules for surrogative reasoning. This contradict our intuitive idea of scientific representation, idea that a reliable theory should account for, rather than contradict. It is exactly this non arbitrary feature of scientific representation that Suárez's objectivity was meant to account for—although I have argued that the inferential theory fails to capture such a notion.

However, an advocate of the interpretational theory could still counter to my objection. At the end of (Contessa's 2007) it is argued that the interpretational theory is not meant to be a full account of scientific representation. It defines epistemic representations, and

> "Although in order to have an account of how models represent, it is necessary to have an account of epistemic representation, in and of itself, an account of epistemic representation does not constitute an account of how models represent. […] The problem is that, even if this surrogative inference is valid according to the set of rules that our interpretation of the model underlies, the inference may well not be sound. The fact that we have adopted an interpretation only implies that the model represents the system and not that it does so faithfully. As I have maintained, the model represents the system faithfully only if all the valid inferences are sound. […] In order to have an account of how models represent their target system, one's account of representation has to be supplemented with an account of how scientific models represent their target systems faithfully. It is only when we have also a solution to this second problem that we will have a full understanding of how models represent their target systems."

This could cast some doubt about the legitimacy of my claim that Contessa's theory leads to paradoxical consequences. Could then an advocate of the interpretational theory appeal to such a distinction in order to solve the apparent contradiction that the interpretational theory implies? I think not. First of all Contessa recognizes that the peculiarity of scientific models with respect to other generic representation does not reside in their truthfulness or accuracy. But if this is so, no eventual further conditions on the truthfulness of the representation can do any work in the definition of scientific representation. Besides this question, however, if taken seriously the above quotation seems to be in open contrast with the declared goals of the inferential theory, i.e. to provide a substantial account of scientific representation. If we were supposed to take seriously the claim that "in and of itself, an account of epistemic representation does not constitute an account of how models represent", we should conclude that, contrarily to Contessa's claims, the inferential theory does not provide a substantial account (i.e. sufficient and necessary conditions) for scientific representation.

4.5.5 Interpretation and surrogative reasoning

Finally, in this section I want to consider a second argument brought by Contessa in favour of his interpretational view and against Suárez's. We have seen that Contessa (rightly, in our view) objects to the inferential view that "the user's ability to perform inferences from a vehicle to a target seems to be a brute fact, which has no deeper explanation. This makes the connection between epistemic representation and valid surrogative reasoning needlessly obscure and the performance of valid surrogative inferences an activity as mysterious and unfathomable as soothsaying or divination." (p.9) An important asset of the interpretational view is therefore that, with the definition of interpretation and the derived notion of valid surrogative reasoning, it explains the nexus between surrogative reasoning and representation.

In arguing in favour of such an explanation Contessa claims that interpretation *is more fundamental* than surrogative reasoning, i.e., that surrogative reasoning occurs *in virtue* of interpretation:

"One performs inferences from the London Underground map to the London Underground network *in virtue of* the fact that the map represents the network. The reverse, however, is not true: the map does *not* represent the network *in virtue of* the fact that one uses it to perform inferences about the network. In fact, we would not even attempt to use a piece of glossy paper with colored lines printed on it to find our way around the London Underground network if we did not already regard the former as an epistemic representation of the latter." (p.14)

However, this 'foundational' role played by interpretation depends strongly on the meaning assigned to the expression "in virtue of". Within Contessa example, the assertion that 'one performs inferences from the London Underground map to the London Underground network *in virtue of* the fact that the map represents the network' can be questioned. As such, the use of such assertion for granting a foundational character to interpretation seems to be a *petitio principii*.

First of all it is to be noticed that, as the kind of interpretation adopted defines the valid surrogative reasoning, in the same way the form that interpretation takes in each instance is strongly affected, in various ways, by the inferences supported by the vehicle. An example of such an influence can be seen for instance in the nowadays well known and recognized role of surplus mathematical structures in the development of scientific theories (French, 1999).

French illustrates in this way the relation of mathematics to physics:

"a theory *T* [is embedded] in a mathematical structure *M'*, in the usual set-theoretic sense of there existing an isomorphism between *T* and a substructure *M* of *M'*. […] An uninterpreted calculus *C* can then be introduced of which *T* and *M* can be taken as isomorphic models, and likewise a calculus *C'* for *M'* can be presented which drives the introduction of a new theory *T'* which in turn is partially interpreted via the structure *T*" (French, 1999, p.188)

It is straightforward to see how this kind of pattern is translatable in Contessa's terminology. A given interpretation I interprets O, the subject of T, in terms of a mathematical structure M. M is a substructure of M', whose additional structure with respect to M allows the derivation of new hypotheses on T. The surrogative reasoning producing such hypotheses leads therefore to a new interpretation I'. Moreover, it is obvious that the composition of the set $\left(\Omega^V = \left\{o_1^V,...o_n^V\right\}\right)$, $\left(P^V = \left\{{}^nR_1^V,...,{}^nR_m^V\right\}\right)$ and $\left(\Phi^V = \left\{{}^nF_1^V,...,{}^nF_m^V\right\}\right)$ of relevant objects, properties and functions of the vehicle, identified by I', can be affected, for instance, by the eventual soundness of the inferences these elements support. The very first interpretation of the planetary model of atom, for instance, originally assumed that the planets' property of always possessing a determinate position was an element P^V. But once the experiments proved that this assumption led to false conclusions, the original interpretation had to be modified, and definite position is now anymore considered as a part of the positive analogy between solar system and atom (Well, besides within Bohmian interpretations of quantum mechanics, of course).

This mutual dependence between surrogative reasoning and interpretation was already recognized by Suárez: "In scientific practice, the requirements expressed in [denotation] and [surrogative reasoning] stand in a dynamical equilibrium. On the one hand the specification of the source and its representational force in part (i) constrains the level of competence and information required of an agent for representation; on the other hand an inquiry into the inferential capacities of A may lead either to shifts in the force of A, or to a reconsideration of what an appropriate source is to represent a given target B." (Suárez, 2004, p. 10).

But there is also another sense in which the interpretation occurs in virtue of surrogative reasoning. It can also be argued, in fact, that the London Underground Map has been made in order to—and therefore has been conceived in the very first place as a mean to—permit informed users to orient themselves inside (i.e. draw conclusions about) the underground network. *Only in second instance, therefore, the map is taken as a representation of the*

underground network. According to this view, therefore, it is *in virtue of its primary function of tool for the orientation* that the underground map is taken to represent the London underground.

An advocate of the interpretational view could therefore argue that the foundational role carried out by interpretation comes from the set of rules provided by interpretation, which grounds validity of surrogative reasoning. The priority of surrogative reasoning in the *intentions* of the user, the argument would continue, is irrelevant, for it is the interpretation that *justifies* surrogative reasoning. Such an objection, however, seems to presuppose a straight (neopositivist flavoured) distinction between *context of justification* and *context of discovery*, and the assumption of the sole relevance, for philosophy of science, of the first. Once it is acknowledged that such a distinction is blurred and that it can be hardly used as a base for the epistemological value of scientific practices, the above objection looses of force.

4.6 Models, representation and surrogative reasoning.

4.6.1 The outline of an account

The discussion I've lead so far has highlighted the role that surrogative reasoning has in scientific representation. The theories I have illustrated in the previous sections have the undeniable merit of recognizing such a role—however, for different reasons, they fail to answer to the basic questions that, following Frigg, we have posed in the beginning of this chapter

Summing up, I have tried to focus the reader's attention towards one precise issue, i.e. the fact that both the inferentialist and the interpretationalist view fail to account for the non arbitrary character of scientific representation, which, so I have argued, should be a constitutive part of the answer to the enigma of representation. I have also argued that Suárez

recognized this non arbitrary feature of scientific representation, and that he have tried to capture it with the notion of objectivity—therefore I also refer to this characteristic as objectivity. The reasons which brought the two theories to failure are different: Suárez clearly put objectivity at the centre of his theory however he fails to grasp the nature of such a notion; Contessa seems to dismiss the problem, but in such a way his theory ends up to be too inclusive.

In section 3.2 I have argued that Suárez fails to capitalize on his intuition of the importance of the context of inquiry when it comes to the characterization of objectivity. First of all, I have argued that, although Suárez claims that the context of inquiry determines additional necessary conditions that representation must meet, he does not explain how and why this is so. Due to this lack of clarity, I argue, the inferential view fails to find out where the objectivity of scientific representation resides. Suárez defines objectivity as informativity and then utilizes surrogative reasoning to characterize it. However, I have shown that the appeal to surrogative reasoning is not sufficient and that, in fact, the way Suárez deal with the proposed examples (the ship-on the sea and the bridge) shows that he in fact implicitly uses the truthfulness of the surrogative reasoning supported by the model as a criterion of demarcation between objective and non objective representation. I have therefore proposed that objectivity should be characterized as a contextual notion, dependent on the kind of activity that the model is supposed to serve.

In the following I will develop this suggestion. More exactly, in the next section I will develop more in detail my definition of objectivity; in sections 4.2 and 4.3, therefore, I will see how objectivity enter in the issues of the problem of style and the enigma of representation, respectively.

4.6.2 Objectivity

In this section I will exactly characterize the objectivity of a scientific representation. We can start with the following definition:

***O**: A scientific model M objectively represents a phenomenon relatively to a given context of inquiry C to the extent to which M supports the achievement of the purposes of C.*

A first obvious consequence of ***O*** is that there is no sense in talking of the objectivity of a scientific representation in the absence of a determinate context of use. In other words, the same model can be an objective representation of a target relatively to a determinate context of use, and contemporarily being a non objective representation of the same target relatively to another context.

An example can be useful to illustrate this point. Take the case of the Kac ring model.

The Kac model has been formulated by Marc Kac ((Kac, 1959), for a simple illustration see Bricmont (1995), Appendix 1) in order to explain Boltzmann's solution to the problem of irreversibility in thermodynamics.

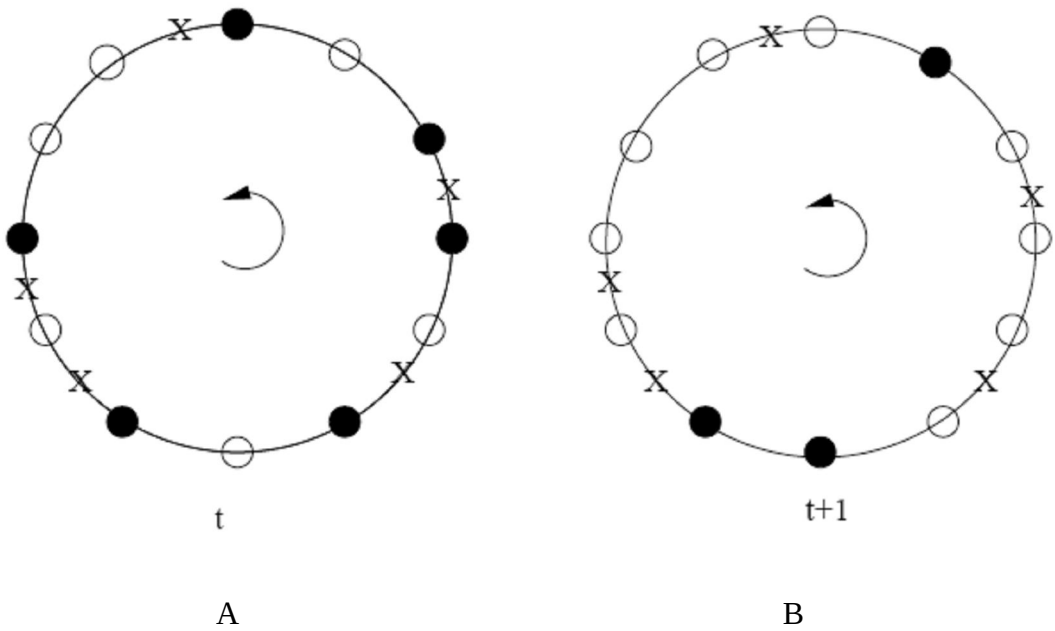

Figure 1: n=12, m=5. A: the ring at instant t. B: the ring at instant t+1. The figure is taken from Roberto Raimondi's webpage: http://www.fis.uniroma3.it/raimondi/iff/lez_11.pdf

Take a ring with *n* equidistant point. At each point is a site containing a ball which can be either white (w) or black (b). *m* intervals between the sites are marked by an *X*. Let's call the set of *m* marked intervals *S* and the set of (*n-m*) non-marked $\underline{S}$. At each elementary time interval, the balls move clockwise, so each ball occupies the nearest site, with the following proviso:

i)the ball remains of the same color if the crossed interval is unmarked;

ii) the ball changes color if the crossed interval is marked by the X.

(see fig.1)

$N_w(t)$ ($N_b(t)$) correspond the number of white (black) balls at time t; $N_w(S; t)$ ($N_b(S; t)$) the number of white (black) balls which are going to cross a marked interval at time t. Within such a model the balls in the ring correspond to the molecules of the gas; the marked intervals correspond to the collision between particles; the colour of the balls corresponds to the (discrete) velocity of the molecules. Finally, the counterclockwise motion of the balls and the rules for the change of the colors denote the equations of motion.

Let's now say that at instant t^0 we have $N_w(t=0)$ white balls and $N_b(t=0)$ black balls. We ask what is the number of white and black balls at time t.

The "equations of motions" of the ring can therefore be written in the following form:

$$N_w(t+1) = N_w(t) - N_w(S;t) + N_w(S;t)$$
$$N_b(t+1) = N_b(t) - N_b(S;t) + N_b(S;t)$$

(1)

Obviously

$$N_b(S;t) + N_w(S;t) = m \qquad (2)$$

Now, in order to use eq. (1), we must know the value of $N_w(S; t)$ and $N_b(S; t)$, and we can do it by checking where the marked intervals are.

Now, we know that the probability *p* for a ball to change the colour is:

$$p = \frac{m}{n} \tag{3}$$

Now, a crucial assumption intervenes in our argument, which is that:

$$\begin{aligned} N_w(S;t) &= pN_w(t) \\ N_b(S;t) &= pN_b(t) \end{aligned} \tag{4}$$[19]

If we then insert (3) and (4) into 1 and then subtract the two resulting equations, we obtain:

$$N_w(t+1) - N_b(t+1) = (1-2p)\left[N_w(t) - N_b(t)\right] \tag{5}$$

Finally, since *(1-2p)<1*, then at each instant the difference between black and white balls decreases.

After a sufficient number *s* of steps, in particular, we expect

$$N_w(t+s) = N_b(t+s) \tag{6}$$

Which, as it is evident, denotes a situation of equilibrium.

The above argument is commonly taken as a simple and effective explanation of irreversibility of thermodynamic processes. It shows that, independently on the state of the ring at the instant t_0 after a sufficient amount of time the final state is very often a state of equilibrium. non equilibrium states are not impossible, but very rare. Finally, the tendency through equilibrium, is shown to rise from the crucial statistic assumption necessary in the passage from the microscopic to the macroscopic description.

The Kac model is widely used for such explanatory function and is obviously also an objective scientific representation of a gas—however, I argue that its objectivity is related to this limited context of use.

To see how this is so, think now of Boltzmann's (1964) introduction to the kinetic theory of gases and, in general, to any illustration to such theories. The most effective model for

[19] For the discussion of the justification of such an assumption, see (Kac, 1959) and (Bricmont, 1995).

picturing how the theory describes gases is here that of the billiard balls, or ideal gas model. In such an illustrating context, under the assumption that the temperature corresponds to the kinetic energy of molecules, we can have, for instance a clear and effective explanation of the Boyle's law. In the present context the Kac model is obviously of no much use. Notice that this is not merely due to the fact that in this model there is no equivalent of the collisions against the container, or for pressure: within Contessa's account, we could always be able to make up an interpretation of the gas in terms of the Kac model in such a way that every relevant element could find a correspondent in the model. The point is instead that the Kac model does not present a causal dynamics which can ground a causal account of the ideal gas law. It is in this sense that, in the context of use of the explanation of irreversibility, the Kac model is an objective representation of gases, but in the context of the explanation of Boyle's law it looses its objectivity.

Notice that it is at this point clear why within the examples used by Suárez (for instance, about the ships on the sea), it was not possible to individuate the difference between an objective and the non objective representation: without specifying the kind of surrogative reasoning which is needed (and which is determined by the context of inquiry), there is no sense in talking of the objectivity of any of the two representations. Let's say therefore that the two pens and the paper are used by two children to play to naval combat (agreed, it is not a scientific practice, but neither two pens and a paper are a scientific model). They must have the possibility of boarding and therefore also to ram the enemy ship, and so on. In this sense, the representation where the pens correspond to the ships and the paper to the sea, it is obviously the more objective, since it supports the naval combat game.

4.6.3 The Problem of Style.

Given the above discussion it should be already obvious how objectivity is crucial for the answer to the problem of style, both in its normative and its actual version. First of all, if

objectivity is a criterion for the normative version of the problem of style, then it follows that the acceptability of a scientific representations have different particular criteria depending on the determinate context of inquiry. It goes without saying that the presented picture also provides a mean for the solution to the factual version of the problem of style. We have for instance a straightforward account of why Suárez's truthfulness or accuracy (corresponding to Contessa's faithfulness) are typically important factors, but anyway not necessary. In many cases, when for instance the theory is too complicated, truthfulness or accuracy end up to be an obstacle for the achievement of an aim. Cartwright (1983, Ch. 8) notoriously argued that the more a model is explanatory, the less it is true. In (Hartmann, 2005) the example of the MIT bag model is illustrated where the model is widely accepted as a good model, due to its exemplificative character and in spite of the evident empirical inadequacy of such models. In such context of inquiry, accuracy is not a necessary condition for a scientific representation, and instead idealization would be required for the simple reason that the former would be an obstacle to the understanding and exploitation of the explanatory relevant factors[20]. The present account of scientific representation therefore also provides a simple account of misrepresentation. As another example, an element which is of crucial importance in the most part of scientific activities is the structure of the model used. It goes without saying that the structural features of the model are essential in practically every use it can be made of it: in the description of the target system—but, for instance, when the model is exploited in order to suggest novel predictions. On the other hand, the fact that models are used for different means (explanation, prediction, description…) does not imply that such kind of scientific investigations are mutually exclusive. A model can be cooked up in order to explain a determinate phenomenon, then it could suggest other developments to the theory, and if it

[20] We are here taking for granted that accuracy is not essential for explanation, however this is another open issues that I cannot deepen here. For a discussion about the role of truthfulness and accuracy in explanation see for instance

ends up to provide novel predictions which are then found to be accurate or true, the model, which now supports also the activity of prediction, is considered more objective.

4.6.4 The enigma of representation.

So far, we have argued that the puzzle of style is accounted for if we say that scientific representations must be objective in the sense that they successfully support some scientific activity. However, we haven't said nothing about the enigma of representation. In (Frigg, 2006) it is rightly argued that the enigma of representation cannot be merely solved by making appeal to the intentionality of the users, since stating that models represent because the agents take them to represent amounts to a mere paraphrase of the problem. An answer to the enigma must instead explain in virtue of what a model is taken to represent a target.

We have seen that within the interpretational theory the enigma of representation is answered by making appeal to interpretation. I have argued that by doing so, Contessa's theory ends up by labelling as scientific paradoxical cases of representation (like in the example of Rutherford's model and the puck on the ice)—and I have concluded that interpretation is therefore not sufficient for scientific representation. For what the inferential theory is concerned—besides the fact that it is not clear what exact role [inf] should play in the solution of Frigg's problems—we have argued that Suárez account failure is due to the fact that it cannot account for the objectivity of scientific representation.

Now, taking objectivity as a criterion for scientific representation amounts to the say first of all that objectivity is a necessary feature for a model to scientifically represent:

R: *A model M is a scientific representation of a target T relatively to a given context of inquiry C only if M successfully supports the achievement of the purposes of C.*

A first important consequence of such a conditional is that not only the criteria of acceptance of a scientific model, but its very relation of "standing for" with respect to the target is context dependent. The discussion of R should not be very problematic. I have

already discussed that a necessary feature of scientific representation is its non arbitrary character. Now, someone could still deny that the non arbitrariness of scientific explanation should be accounted in terms of the contextual notion of objectivity that I have proposed in O. However, I have shown in great detail in the previous chapters how the possibility of capturing the non arbitrariness of scientific representation in non contextual terms is at least very dubious. This should be enough to demonstrate that objectivity is in fact a necessary condition for scientific representation.

Is objectivity also sufficient for scientific representation? If objectivity would be sufficient for representation, it should be possible to show that only provided objectivity, an agent would be lead to take a determinate model to refer to a determinate agent for representation. Remember now the discussion I've carried out in §3.6, about Contessa's claim that it is in virtue of interpretation that a model is capable to support surrogative reasoning. In arguing against that claim I have shown that in many instances the situation is exactly the opposite: the London Underground map is *previously* conceived as a mean for a user to reach one metro station starting by another metro station, and it is typically in terms of such a use that we interpret the map as a representation of the London Underground. But scientific models are typically conceived in these terms. They are never representation *period*, unrelated to any specific use. To be sure, once that a model is used and taken as a representation of a target it can be applied to different means, and his eventual success in different applications straighten his representative power. But the point is that scientific models are always formulated and proposed for a scope, and it is in virtue of the fact that this scope is achieved that the referential capacity of a model is then decided.

4.6.5 Conclusions

The account proposed in the previous section offer a very simple picture of the relation between explanation and representation, and very straight answer to the objections about what kind of knowledge does an explanation like structural explanation provide. My point of view is that the representation provided by the model does not constrict the kind of understanding that a structural explanation provides—but, contrarily, it is the use of a mathematical model for a structural explanation that determines the fact that such a model represent the phenomenon. Agreed, what I have said is far from being a complete account of scientific representation and models. Much more should be said in order to fully understand how do models come to refer, and to spell out in detail an account of all the styles of scientific representation.

The idea proposed in the previous section may suggests a somewhat antirealistic attitude towards representation. To be sure, it is not compelled to any realistic assumption, as a specific objective relation between the model and the target system. However, I think that a realist point of view can anyway be compatible with the above account. I consider in fact an asset of the illustrated proposal the fact that it does not presupposes neither realism or antirealism and that it is therefore compatible with both positions. To be sure, there are some situations in which realistic commitments about the model are undeniably strong—evidently, the fact that a surplus structure of the model can be used in order to formulate novel predictions or in order to expand the old theory represent a strong argument for the truthfulness of (part of) the structural properties of the model.

But the arguments for the necessity of such a realistic commitment should always be grounded on the practice supported by the model (as it was the case of when a prediction is based on the structure of the model), not on any argument on the representative function of the model.

Chapter 5

Explanation in Quantum Information Theory

5.1 Introduction.

It is often argued against Jeffrey Bub's analysis of the philosophical meaning of the Clifton Bub Halvorson Characterization Theorem (therefore CBH) (Clifton, Bub and Halvorson (2003), Bub (2004) and (2005)) that as an axiomatisation of the formalism of quantum mechanics, the CBH theorem cannot represent more than a convenient tool for guiding the research on the foundations of quantum mechanics. As a consequence, this kind of criticisms leads to the claim that Bub's formulation of quantum theory as a principle theory, based on the three information theoretic constraints of the CBH theorem, cannot genuinely explain quantum phenomena.

On the other hand, by exploiting the parallel between QIT and Einstein formulation of Special Relativity (SR), Bub argues that as SR made the Lorentz' theory explanatorily superfluous, so does QIT with respect to other constructive theories. Sometimes explicitly,

sometimes more implicitly, Bub's defence of QIT often involves some appeal to the greater explanatory capacity of QIT with respect to Bohm's interpretation of QM?, other times he more modestly claims that "the lesson of modern physics is that a principle theory is the best one can hope to achieve as an explanatory account of quantum phenomena" (Bub, 2005, p. 15). However, given that the analogy he draws between SR and QIT is extended also to the explanatory level, the logical alternatives are two: either both SR and QIT must be considered as a small consolation after the failure of those theories (Lorentz' and Bohm's) that were able to provide a real understanding of phenomena—**or** both SR and QIT explain phenomena as well, or better, than their rival theory.

While he often challenges the explanatory power of constructive theories à la Bohm, Bub never in effect says *what kind* of explanations or understanding of quantum phenomena we can hope to gain from his principle reconstruction of quantum theory.

In this chapter we will analyze Bub's version of QIT and argue that within it this, quantum theory is meant to provide structural explanations of phenomena. Moreover, we will consider Bub's parallel between SR and QIT at the explanatory level, and try to understand if it is true that if we take one (SR) to provide a good structural explanation, so we should do with the other.

5.2 Quantum Information Theory.

Bub's interpretation takes quantum mechanics as a *principle theory*[21] about the possibility and impossibility of information transfer. Here for information it is meant information in the physical sense, i.e. Shannon entropy: a measure of the uncertainty associated with a random variable. It quantifies the information contained in a message, usually in bits or bits/symbol. It

[21] For the definition of principle theories and the distinction between principle and constructive theories see §2.1.

is the minimum message length necessary to communicate information. Equivalently, Shannon entropy is defined as the amount of classical information we gain, on average, when we learn the value of a random variable.

In section 2.1, we have seen how Special Relativity (SR) was defined by Einstein as a principle theory. SR was born in opposition to Lorentz's constructive theory of the electromagnetic field and with the aim to avoid all the problems it presented. In the same way, Bub interprets quantum theory as a principle theory about the possibility and impossibility of information transfer, opposed to an interpretation of quantum theory as a constructive theory about the behaviour of non-classical waves or particles and based on 3 information-theoretic constraints:

1) **No superluminal information transfer via measurement**. This constraint states that merely performing a local (non-selective)[22] operation on a system A cannot convey any information to a physically distinct system. This constraint corresponds to the no-signalling via entanglement featuring in ordinary quantum mechanics.

2) **No broadcasting**. States the impossibility of perfectly broadcasting the information contained in an unknown physical state. Broadcasting is a generalization of the process of cloning. This, in turn, is a process that starts with a system in any arbitrary state $|\alpha\rangle$ and ends up with two systems, each in the state $|\alpha\rangle$. While cloning applies only to pure states, broadcasting generalizes also to mixed states, and, for pure states, reduces to cloning. In quantum mechanics, broadcasting is possible for a set of states ρ_i iff they are commuting. (Barnum et al. 1996)

[22] Selective measurements operations are here obviously not considered, given that in such operations statistics in general changes given that the selection changes the ensemble under study.

3) **no bit commitment**. The bit commitment is a cryptographic protocol in which one party, Alice, supplies an encoded bit to a second party, Bob, as a warrant for her commitment to the value 0 or 1. The information available in the encoding should be insufficient for Bob to ascertain the value of the bit at the initial commitment stage, but sufficient, together with further information supplied by Alice at a later stage (the 'revelation stage') when she is supposed to open the commitment by revealing the value of the bit, for Bob to be convinced that the protocol does not allow Alice to cheat by encoding the bit in a way that leaves her free to reveal either 0 or 1 at will. As an illustration of how this cheating strategy should work, take this example from (Timpson, 2004):

"Consider a spin-1/2 system: a 50/50 mixture of spin-up and spin-down in the z-direction is indistinguishable from a 50/50 mixture of spin-up and spin-down in the x-direction—both give rise to the maximally mixed density operator ½ 1. Alice might associate the first type of preparation with a 0 commitment and the second with a 1 commitment. Bob, when presented with a system thus prepared will not be able to determine which procedure was used. Alice also needs to keep a record of which preparation procedure she employed, though, to form part of the evidence with which she will convince Bob of her probity at the revelation stage. Thus, for a 0 commitment, Alice could prepare a classically correlated state of the form:

$$\text{0 commitment: } \rho_0 = \frac{1}{2}\left(\left|\uparrow_z\right\rangle_1\left\langle\uparrow_z\right|_1 \otimes \left|\uparrow_z\right\rangle_2\left\langle\uparrow_z\right|_2 + \left|\downarrow_z\right\rangle_1\left\langle\downarrow_z\right|_1 \otimes \left|\downarrow_z\right\rangle_2\left\langle\downarrow_z\right|_2\right)$$

whilst for a 1 commitment, she could prepare a state

$$\text{1 commitment: } \rho_1 = \frac{1}{2}\left(\left|\uparrow_x\right\rangle_1\left\langle\uparrow_x\right|_1 \otimes \left|\uparrow_x\right\rangle_2\left\langle\uparrow_x\right|_2 + \left|\downarrow_x\right\rangle_1\left\langle\downarrow_x\right|_1 \otimes \left|\downarrow_x\right\rangle_2\left\langle\downarrow_x\right|_2\right)$$

System 2 is then sent to Bob.

At the revelation stage, Alice declares which bit value she committed to, and hence which preparation procedure she used. The protocol then proceeds in the following

way: If she committed to 0, Alice and Bob both perform σ_z measurements and Alice declares the result she obtains, which should be perfectly correlated with Bob's result, if she really did prepare state ρ_0. Similarly, if she committed to 1, Alice and Bob both perform σ_x measurements and Alice declares her result, which again should be perfectly correlated with Bob's result, if in truth she did prepare state ρ_1. If the results reported by Alice and obtained by Bob don't correlate then Bob knows that Alice is trying to mislead him. The trouble with this otherwise attractive protocol is that Alice is able to cheat freely by making use of what is known as an EPR cheating strategy. Thus, rather than preparing one of the states ρ_0 or ρ_1 at the commitment stage, Alice can instead prepare an entangled state, such as the Bell state $|\varphi^+>_{12}$. The reduced density operator for Bob's system will still be ½1, but Alice can now simply wait until the revelation stage to perform a suitable measurement on her half of the entangled pair and prepare Bob's system at a distance in whichever of the two different mixtures she chooses" (pp. 212-213)

By asserting the impossibility of such a secure cryptographic protocol, the no-bit commitment principle assures the stability of entangled states also in macroscopic or nonlocal processes. In other words, the no-bit commitment in incompatible with theories which are like quantum mechanics locally, in the sense that they allow the existence of ambiguous mixtures, but where entangled states decay in macroscopic or nonlocal states (Schrödinger contemplated the possibility of such a theory in (Schrödinger, 1936)). Following these theories (which, following Timpson, we will call *Schrödinger-type theories*), the EPR cheating strategy would not be applicable (given that entangled states would not be stable enough) and the secure bit-commitment in general possible.

The CBH Characterization Theorem, therefore, demonstrates that the basic *kinematic* features of a quantum-theoretic description of physical systems (i.e. noncommutativity and entanglement) can be derived from the three information-theoretic constraints.

The formal model utilized by QIT in order to derive such a result is the C*-algebra. This is an abstract representation of the algebra of observables which can characterize both classical (mechanical particle and field) and quantum mechanical theories[23]. More precisely, every *commutative* C*-algebra corresponds to a classical phase theory, and to every classical phase space theory corresponds a C*-algebra. On the other hand, CBH characterizes quantum mechanical theories as *noncommutative*, but mutually commuting, C*-algebras.

It is important to note that that of a C*-algebra is a not a necessary nor obvious choice, for weaker algebras (for instance Segal algebras) could also be adapt to characterize a quantum theory. Moreover, the existence of 'toy theories' that, although satisfying the CBH constraints are not quantum mechanical theories can raise some doubts on the choice of C*-algebra as a mathematical framework.[24]

For what quantum mechanics is concerned, the algebra $\mathsf{B}(\mathsf{H})$ of all bounded operators on a Hilbert space H is a C*-algebra, with * the adjoint operation and $\|\cdot\|$ the standard operator norm. A state on a C*-algebra C is defined as any positive normalized linear functional $\rho : \mathfrak{C} \rightarrow \mathbb{C}$ on the algebra. A state is pure iff when $\rho = \lambda\rho_1 + (1-\lambda)\rho_2$ with $\lambda \in (0,1)$, then $\rho = \rho_1 = \rho_2$. Pure states of $\mathsf{B}(\mathsf{H})$ are admitted that are not representable by vectors in H (nor by density operators in H). A representation of a C*-algebra C is any mapping $\pi : \mathsf{C} \rightarrow \mathsf{B}(\mathsf{H})$ that preserves the linear product and the linear $*$ structure of C.

[23] For a more detailed illustration of the C*-algebraic formalism see Bub (2004, 2005) and Timpson (2004)
[24] See Timpson 2004, Chapter 9.

A quantum system A is represented by a C*-algebra **A** and a composite system A+B is represented by the C*-algebra **A** v **B**. Observables are represented by self-adjont elements of the algebra. A quantum state is an expectation-valued functional over these observables. The constraint is added that two systems A and B are physically distinct when any state of **A** is compatible with any state of **B** (C*-independence), that is, for any state ρ_1 of **A** and for any state ρ_2 of **B**, there is some joint state ρ of the joint algebra **A** v **B** such that $\rho|_A = \rho_1$ and $\rho|_B = \rho$.

Having said this, the CBH theorem demonstrates how quantum theory (which, again, they take to be a theory formulated in C*-algebraic terms for which the algebras of observables pertaining to distinct systems commute, for which the algebra of observables on an individual system is noncommutative, and which allows space-like separated systems to be in entangled states) can be derived from the assumption of the three information-theoretic constraints. More exactly, it is demonstrated that: 1) from the first constraint, no superluminal information transfer via measurement, it follows that commutativity of distinct algebras is guaranteed (the converse result is proved in (Halvorson, 2003): if the observables of distinct algebras commute, then the no superluminal information transfer via measurement constraint holds). Commutativity of distinct algebras is meant to represent no-signalling; 2) CBH demonstrate both that cloning is always allowed by classical (i.e. commutative) theories and that, if any two states can be (perfectly) broadcast, then the algebra is commutative. Therefore, from the second constraint, no broadcasting, follows the noncommutativity of individual algebras. Noncommutativity of individual algebras is the formal representative of the physical phenomenon of interference. 3) if **A** and **B** represent two quantum systems (i.e., if they are noncommutative and mutually commuting), there are nonlocal entangled states on the C*-algebra **A** v **B** they generate. This result has been reached in some works by (Landau, 1987, Summers, 1990, Bacciagaluppi, 1993).

However, Bub argues, we still cannot identify quantum theories with the class of noncommutative C*-algebras. It is at this point that the third information-theoretic constraint, the no unconditionally secure bit-commitment, is introduced, 'to guarantee entanglement maintenance over distance'.

Timpson notices that the role of no bit-commitment in this sense somewhat ambiguous.

The firstly suggested motivation for the need of the no bit commitment is in fact the following: the arising of nonlocal entangled states in the account so far provided, follows directly from the choice of the C*-algebra and from its formal properties. On the other hand, "in an information-theoretic characterization of quantum theory, the fact that entangled states can be instantiated nonlocally, should be shown to follow from some information-theoretic principle."(Bub, 2004, p.6). It seems, in other words, that the role of the no bit-commitment is to provide an information theoretical ground *in the context of C*-algebra* to the rising of entanglement, which, otherwise, would be a consequence of the sole mathematical machinery used by the theory. This suggestion is made more clear in (Clifton, Bub and Halvorson, 2003):

> So, at least mathematically, the presence of nonlocal entangled states in the formalism is guaranteed, once we know that the algebras of observables are nonabelian. What does not follow is that these states actually occur in nature. For example, even though Hilbert space quantum mechanics allows for paraparticle states, such states are not observed in nature. In terms of our program, in order to show that entangled states are actually instantiated, and—contra Schrödinger—instantiated nonlocally, we need to derive this from some informationtheoretic principle. This is the role of the 'no bit commitment' constraint. (p.10)

But if the mathematical structure of reference is a C*-algebra, it would seem that the function of the third principle would be to *reassess* the occurrence of entangled states, which

are already part of the theory. But the idea of positing a principle in order to "rule in" something which is already part of the theory is pretty peculiar: "ruling states *in* rather than *out* by axiom seems a funny game. Indeed, once we start thinking that some states may need to be ruled *in* by axiom then where would it all end? Perhaps we would ultimately need a separate axiom to rule in every state, and that can't be right." (Timpson, 2004, p.206) On the other hand, given that the problem seems to rise from the existence of other weaker algebras where entanglement could not follow from the first two principles, the no-bit commitment could be seen as a constraint on this more general context. But in this case, it is still to be proved that no bit-commitment would succeed, given that so far there is no proof that it would guarantee in this more general context the stability of non-local entanglement.

Other times Bub suggests that the function of the no bit commitment is slightly different. We have already seen that the no-bit commitment is incompatible with Schrödinger-type theories that, even if not in violation of the no information via measurement and no broadcasting principles, eliminate non local entanglement by assuming, for instance, its decay with distance. Timpson argues that also this argument is anyway dubious, since "a Schrödinger-type theory is only an option in the sense that we could arrive at such a theory by imposing further requirements to eliminate the entangled states that would otherwise occur naturally in the theory's state space." (Timpson, 2004, p.207)

In (Hemmo and Hagar, 2006, n.12 and 19) the no-bit commitment is interpreted as a *dynamical* constraint, meant to rule out dynamical theories (such as GRW), which, still coherent with the first two principles, implies a decay of entanglement at the macroscopic level. Timpson also considers this option (2004, Ch. 9) but, rightly in our view, rapidly rejects it as in evident contrast with Bub's theory's manifested ambitions of being concerned on the "kinematic features of a quantum-theoretic description of physical systems" (Bub, 2004, p.1). But let's anyway explore where this interpretation would lead. As noticed by Hagar and Hemmo themselves, also in this case the no bit commitment maintain a controversial status.

Recall the previous illustration of the bit commitment procedure. In standard quantum mechanics the no bit commitment holds since entanglement is stable also at a distance, therefore Alice can always cheat by sending to Bob a particle in entangled state. Given the well known result of the Aspect experiment, we know that in such a situation a Schroedinger type theory postulating a decay of the entangled state is not empirically adequate. This is the reason why, in such a case, the no bit commitment is justified. On the other hand, in this kind of situation the no bit commitment is respected also by the GRW theory, since the entangled state is stable at microscopic scale (also when the particles are far). GRW violates the no bit commitment just in case the entangled state concerns a massive system, since in this case the entanglement decays very quickly. But in this case, also standard quantum mechanics implies an *effective* decay of the system (and therefore an *effective* violation of the no bit commitment), due to decoherence. And as for the moment there is no empirical ground for deciding which one of the two prediction (GRW's collapsed state or standard quantum theory decohered state) is true. But then it follows that if the no bit commitment is meant to ensure the stability of entanglement also at a distance, then it is uninformative; if it is meant to ensure the stability of entanglement also for massive bodies, then it is not supported by empirical grounds. (Agar and Hemmo, 2006, §3.2)

But there is another weird consequence that should convince to abandon the interpretation of the no bit commitment as a dynamical principle. So far we have utilized Agar and Hemmo's treatment of the no bit commitment in order to show how, even if taken as a dynamical principle, it is not able to provide an information-theoretic ground to the occurrence of entangled states. Now, consider again the option depicted above of the no bit commitment having an active role in ruling out situations of decay of the entangled states in the GRW theory. To rule out this option would mean to forbid not only collapse in non local entangled states, but collapse in massive bodies. In other words, what we would end up with

would be a no-collapse theory. But this is not clearly what Clifton, Bub and Halvorson had in mind with the third information-theoretic principle.

Summing up, these are the conclusions we reached with respect to the effectiveness of the no bit principle in providing an information-theoretic ground to entanglement: as a kinematic principle, the no bit commitment has a dubious role: either it is redundant (in the context of the C*-algebra); or it is unconvincing (in the case of Segal algebra). As a dynamical principle…also: either it merely applies to cryptographic procedures where the entangled states utilized are states of microsystems, in which case is still redundant (since also the GRW respect it); or it also applies to situations where massive systems are concerned, in which case it would be unfounded, and it would make the QIT correspond to a no-collapse theory.

We will come back to the role of the no-bit commitment later on, since it will have a crucial role in the discussion of structural explanation in QIT.

To conclude with our exposal of QIT, these are the fundamental three theses defended by Bub on the significance of the CBH theorem:

- *A quantum theory is best understood as a theory about the possibilities and impossibilities of information transfer, as opposed to a theory about the mechanics of nonclassical waves or particles.*
- *Given the information-theoretic constraints, any mechanical theory of quantum phenomena that includes an account of the measuring instruments that reveal these phenomena must be empirically equivalent to a quantum theory.*
- *Assuming the information-theoretic constraints are in fact satisfied in our world, no mechanical theory of quantum phenomena that includes an account of measurement interactions can be acceptable, and the appropriate aim of physics at the fundamental level then becomes the representation and manipulation of information.*

5.3 The problem of explanation in QIT

In addition to the above illustrated three principal thesis, Bub also often opposes QIT's explanations of quantum phenomena to those provided by constructive interpretations of quantum mechanics—suggesting some times that QIT's explanations are as satisfactory as any other constructive quantum theory, other times advancing the more modest claim that "the lesson of modern physics is that a principle theory is the best one can hope to achieve as an explanatory account of quantum phenomena" (Bub, 2005, p.19).

Against Bub's claims, however, it could be argued that the point is not whether QIT's explanations are more or less acceptable than, say, Bohm's theory's explanations, but whether QIT can provide any explanation at all.

This problem rises obviously if one is committed to a 'constructive' view of scientific explanation. More precisely, someone endorsing Harvey Brown's analysis of explanation within SR would most likely question the parallel proposed by Bub between SR and QIT. Remember (chapter---) that following Brown, within the current "orthodox" conception of SR, spacetime is considered as an entity "of a special kind" which, so to say, shapes lengths and time (Brown and Timpson, p. 11, or Brown and Pooley, 2004, p.14). This, Brown's argument continues, provides a "constructive" dimension to Minkowski's formulation of special relativity and, at least in the intentions of the proponent of the orthodox view, makes the geometry of spacetime explicative with respect to the relativistic effects.

Now, this kind of analysis (which, by the way, we don't find convincing anyway) is obviously not applicable to QIT. Even if he does acknowledge the status of a primitive physical quantity to information, Bub clearly rejects a view of QIT as providing a "constructive explanation" of quantum phenomena, with the structure of information acting as a sort of 'cause' of the occurrence of quantum phenomena.

But if not a constructive kind, what kind of explanations or understanding of quantum phenomena can we hope to gain from QIT?

Bub claims that, given the CBH theorem, QIT makes Bohm's theory explanatorily irrelevant. On the other hand, in order to say this, Bub must assume that what is explained by Bohm's theory is already explained by QIT. Bub's argument is that:

> "if the information-theoretic constraints apply at the phenomenal level, then, according to Bohm's theory the universe must be in the equilibrium state, and in that case there can be no phenomena that are not part of the empirical content of a quantum theory (i.e., the statistics of quantum superpositions and entangled states)."

From which it follows that:

> "the additional non-quantum structural elements that [no collapse hidden variable] theories postulate cannot be doing any work in providing a physical explanation of quantum phenomena that is not already provided by an empirically equivalent quantum theory" (Bub (2005), p.12)

However, this argument is dubious. Put in other (much less elegant!!) words, it would mean: given that QIT predicts the very same phenomena as Bohm's theory, then it also equally explains them. Moreover, in order to be consistent, the above argument must presuppose that empirical prediction is a sufficient condition for explanation—it also seem to imply that no other factor contributes to the explanatory power of a theory. However, with no further assumption on what it is to be counted as an explanation, there seems to be no reason for taking the prediction of, say, entanglement as a sufficient condition for its explanation (let alone for the only criterion for its explanation). If this is true, therefore, Bub's argument can

only be defended within the context of a Deductive-Nomological view of explanation[25], where explanation and prediction are intimately linked. An information-theoretic explanation of the existence of entangled states would therefore consist in a logical derivation of the sentence "a quantum entangled states occurred" from the sentences expressing initial conditions and laws of nature, where the laws of nature would be the three information-theoretic constraints.

This hypothesis, however, would obviously weaken Bub's position, for the acceptability of the latter would be bound to the acceptance of the DN as a satisfactory model of explanation. In other words, if these are the kinds of explanations that we can derive from QIT, it is reasonable to object that a pure logical derivation does not seem to help much our understanding of quantum entanglement.

Moreover, given what have been saying before on the role of the no bit commitment and of entanglement, there are good reasons to think that also a DN explanation would not be realizable of entanglement. The obvious candidates for acting as the laws of nature in QIT, in fact, are the three information-theoretic principle, but we have already seen how the CBH theorem does not provide a full information-theoretic account of entanglement, given a) the arbitrary choice of C*-algebra over other admissible algebras, b) the fact that there is no proof of the derivability of entanglement from the first two information-theoretical in the context of weaker algebras, c) the introduction of the no bit-commitment is useless to solve the problem.

In other words, entangled states are not logically deducible from the three information theoretic principles.

[25] If for Bub, as it seems, the special character of SR's explanations lies in the different, "principle", method for the inference of the explanandum, then, given the "natural" place that the DN model occupies within an 'axiomatic' view of scientific theories, the DN model seems to be the natural candidate for accounting for the explanatory power of STR

5.4 Structural explanation.

A more convincing way to go, therefore, could be offered by structural explanation: we explain entanglement with QIT by showing how entanglement is part and parcel of the formal model displayed by QIT (i.e. noncommutative C*-algebra), what its role is in the formalism and its relations with other explicit features of such a formalism.

Moreover, the fact that entanglement rises in QIT as a consequence of the mathematical properties of C*-algebra should not represent a problem here, since (in our view) is in the very nature of structural explanation the fact that it exploits the mathematical resources of the theory.

This interpretation is especially suggested by Bub's parallel between QIT and SR, which explicitly applies also to the explanatory level:

> "I argue that just as Einstein's analysis (based on the assumption that we live in a world in which natural processes are subject to certain constraints specified by the principles of special relativity) shows that the mechanical structures in Lorentz's constructive theory (the ether, and the behaviour of electrons in the ether) are irrelevant to a physical explanation of electromagnetic phenomena, so the CBH analysis (based on the assumption that we live in a world in which there are certain constraints on the acquisition, representation, and communication of information) shows that the mechanical structures in Bohm's constructive theory (the guiding field, the behaviour of particles in the guiding field) are irrelevant to a physical explanation of quantum phenomena." (Bub, 2004)

In Ch. (?) it has been argued that SR's explanations of relativistic phenomena are structural explanations—from Bub's account, therefore, it should follow that the same applies to QIT. In this section we therefore want to put forward this hypothesis and discuss to which extent, under this light, Bub's parallel between SR's and QIT's explanations holds.

Let's see more in details how this kind of account should explain entanglement. We have seen that the CBH theorem starts with the choice of C*-algebra as the background mathematical structure, and how this algebra covers various different physical theories, both classical and quantum. We must therefore first notice how within the framework of C*-algebras classical theories are differentiated by quantum theories by the fact that while the former are characterized by commutative C*-algebras, the C*-algebras representing the latter are non-commutative. This difference is crucial for a structural understanding of entanglement within the context of the CBH theorem, given that, as we have seen above (section 1) it can be shown that if **A** and **B** are two noncommutative and mutually commuting C* algebras, there are non local entangled states on the C*algebra **A** v **B** they generate: "[s]o it seems that entanglement—what Schrödinger (1935, p. 555) called '*the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought'—follows automatically in any theory with a noncommutative algebra of observables. That is, it seems that once we assume 'no superluminal information transfer via measurement', and 'no broadcasting,' the class of allowable physical theories is restricted to those theories in which physical systems manifest both interference *and* nonlocal entanglement." (Bub, 2004, p. 6)

At first sight, this would already constitute a structural explanation of the entanglement: Bub's argument has shown how entanglement is part and parcel of the formal structure of any quantum theory, i.e. any theory characterized by a noncommutative C*-algebra. Not only: the effectiveness of this structural explanation comes from the fact that it highlights how entanglement raises within quantum theories from the noncommutative character of their structure, and therefore, why it does not occur in classical (viz. commutative) theories. In

other words, this structural explanation highlights the necessary relation of entanglement with other explicit elements of the formal structure.

The question, however, is not so simple. We have said that the fact of exploiting the mathematical properties of C*-algebra does not represent a problem for structural explanation. However, the problem still remains of the availability of other algebras where possibly entanglement would not follow. If in our version of structural explanation it is admitted in (and it is indeed congenital of) structural explanation to use structures that are essentially mathematical, this does not imply that any mathematical model can do the work. More exactly, in a situation like the present, where two different models (C*-algebra and Segal algebra) seems to be acceptable, and the explanandum is not an element of one of them, a structural account of the explanandum in terms of only one of the models is obviously to be considered partial, at least. We are not asking here for an information-theoretic derivation of entanglement. But just for a straight derivation.

To sum up, here's what our analysis came up with. Following Bub, QIT is not meant to provide a 'constructive' or causal explanation of quantum phenomena—however, against Brown, we argued that this alone does not imply that QIT lacks explanatory power. On the other hand, not even a DN explanation of entanglement seems realizable within Bub's theory: a satisfactory (information-theoretic) DN explanation of entanglement needs either an argument (relying on information-theoretic bases) which compels to the adoption of a C*-algebra, or an argument (relying on information-theoretic bases) which can assure the rising of entangled states also in weaker algebras than the C*-algebra. Following Timpson's analysis and against Bub's suggestion, we have argued that it is unsure that the no-bit commitment could effectively work as the needed information theoretic base for these two options

We therefore advanced the idea that QIT is aimed to provide structural explanations: a quantum theory is characterized by a non-commutative C*-algebra, and we understand

entanglement in the context of the CBH theorem as a basic feature of any non-commutative C*-algebra. But we have seen how also as a structural account, QIT can provide at best a partial explanation.

We are now in the condition to reconsider Bub's claim that as STR makes Lorentz's theory explanatorily irrelevant, so does QIT with respect to the constructive interpretations of quantum mechanics. For what we have argued so far there is a big difference between the structural explanation that STR and QIT in Bub's version. In our account of structural explanation within STR, we have illustrated how, if not impossible, a constructive dynamical explanation of relativistic effects is not needed in order to fully understand relativistic effects. The same, we have argued, cannot be said about QIT, which is far from providing a satisfactory structural account of entanglement (let alone information-theoretic). As long as such an account is not completed, QIT will not be able (borrowing the term from Wesley Salmon) to 'screen-off' constructive interpretations of quantum mechanics, as SR does with Lorentz' constructive theory.

5.5 Explanation and interpretation.

In this section we finally consider the consequences that our analysis of explanation in QIT can have in the debate on the interpretation of quantum mechanics. Even if never talked out in details, the question of the explanatory power of QIT often raises its head in many criticisms moved against QIT as fundamental interpretation of quantum mechanics.

As an example of such an argument take Brown and Timpson's (2006) discussion of QIT. The target of such a discussion is Bub's philosophical interpretation of SR as a principle theory and the application of the latter to the case of quantum mechanics.

In (Bub, 2006) it is presented a historical fable, where "the special theory of relativity was first formulated geometrically by Minkowski rather that Einstein, as an algorithm for relativistic kinematics and the Lorentz transformation, which is incompatible with the kinematics of Newtonian space-time." (Bub, 2006, p.5) Without Einstein reformulation of special relativity as a principle theory, Bub goes on, Minkowski's relativity would have surely been considered nothing more than a "convenient (but 'counterintuitive and mind-boggling') algorithm". Bub's conclusion is therefore that, as for Einstein's analysis with respect to Minkowski's formulation of SR, the CBH characterization theorem provides us with the rationale for taking quantum mechanics as a principle theory.

Let's now pass to Brown and Timpson's objection. We have already mentioned in §3 of this chapter that in Brown's view, the orthodox interpretation of SR conceives Minkowski's spacetime as the holder of the explanatory burden with respect to relativistic effects and, consequently, it is interpreted as an "entity of a special kind".. It is this "constructive dimension" of STR that supplies (at least in the intention of the supporters of the "new orthodoxy") explanatory force to the theory, and it is for this reason that Minkowski's spacetime can in no way, be reduced to a simple algorithm. It is evident how this argument suggests the subsistence of an inevitable contrast between the interpretation of STR as a principle theory (and of Minkowski's spacetime as a mere mathematical background for its formulation) and its explicative capacity. It goes without saying that according to Brown and Timpson, this contrast would also apply to quantum mechanics, whose interpretation as a principle theory would necessarily disadvantage its explanatory capacity.

Now, under the light of what we've said in the previous section and in Ch. (----) about structural explanation in QIT and STR, we wonder whether Brown and Timpson's objection still holds. If our analysis of structural explanation in STR is on the right track, this means that in order for the geometry of space-time to explain relativistic effects, *it is not necessary to assume space-time to be a special kind of entity providing a "constructive" dimension to*

STR's explanations of relativistic phenomena. Therefore, if we are right, the undeniable explanatory capacity of STR is in no way an evidence against its interpretation as a principle theory—let alone against the interpretation of quantum mechanics as a principle theory.

In his seminal paper about structural explanation in quantum mechanics, Robert Clifton suggests that choosing the algebraic approach to quantum mechanics, should not imply to take a realist stance about the algebraic structure of observables. The program should be instead to "capture the intrinsic structure of relativistic quantum field theory by associating algebras of local observables with regions of spacetime. In the 'concrete' approach to the theory, those observables are constructed out of quantum fields, but once constructed the algebraic approach counsels us to throw that ladder away" (Clifton, 2001, p. 19).

In the same way, we don't need to take a realist stance towards information in order to appreciate the insight that QIT, and the CBH theorem in particular, offer of the quantum world.

We have seen that QIT is far from supplying a satisfactory structural explanation of quantum phenomena, and it is still unsure if it will ever provide it.

But we suggest that to work in this direction and acknowledge the potential explanatory power than QIT would have with structural explanation, one should first of all reject the often implicitly imposed dichotomy: constructive-fundamental-explicative theories Vs principle-phenomenological-non explicative theories. Secondly, the difficulties risen in providing a fully informational-theoretic account of quantum theory should suggest that the right 'epistemological stance' to take towards QIT would more likely be to exploit its mathematical resources in order to discover the intrinsic structure of quantum theory, and then throw away the ladder (hoping to gain with it a structural explanation of quantum phenomena)—rather than to interpret information as a new physical primitive and QIT as a fundamental theory.

Chapter 6

Nonlocality in the Many Minds Interpretation

6.1 Introduction.

In this paper we shall discuss an argument proposed by Meir Hemmo and Itamar Pitowsky (Hemmo and Pitowsky, 2003) on the nonlocality of the Many Minds Interpretation (MMI, hereafter) in the version given by David Albert and Barry Loewer (Albert and Loewer, 1988, and Albert, 1992). Hemmo and Pitowsky claim that even if Albert and Loewer's theory is not strongly nonlocal in Bell's sense, it still exhibits a *weak form* of nonlocality. Their proof for this can be articulated in three main points that we want to discuss: i) in order to be coherent with Albert and Loewer's position on the so-called *mindless hulks problem* it is necessary to admit *weak minds-correlations* between (sets of) minds

of different conscious beings; ii) the assumptions of transtemporal identity of minds and of weak minds-correlations together entail a dependence of the partition of minds of one observer on the measurements performed by the other observers (this dependence is supposedly shown in Hemmo and Pitowsky's analysis of the GHZ experiment); iii) these correlations are (weakly) nonlocal.

An important part of our discussion will be devoted to a criticism of Hemmo and Pitowsky's analysis of the mindless hulks problem, which plays a crucial part in step i) of their argument. *Contra* Hemmo and Pitowsky, we will argue that the core of the mindless hulks problem is the violation of the supervenience of the mental states on the physical states, and we will try to show how this problem is pivotal for the final shape Albert and Loewer gave to their theory. The main part of our criticism, however, will be directed against Hemmo and Pitowsky's attempt to derive ii).

We shall start by presenting a brief review of the original motivations for the MMI (section 2). Such a review will show how the mindless hulks problem is the product of a tension between, on the one hand, the need for solving the determinism problem, on the other hand the need for rescuing the supervenience of the mental on the physical. In section 3, we shall present Albert and Loewer's argument for the locality of their theory. In section 4, we shall discuss Hemmo and Pitowsky's argument. First (section 4.1) we shall define in detail in what the weak minds-correlations consist. Then (section 4.2) we shall argue, against ii), that the dependence of the partition of minds of one observer on the measurements performed by the others (say, in a GHZ experiment) cannot be deduced from the weak minds-correlations alone. Instead, Hemmo and Pitowsky's argument presupposes a stronger assumption of minds-correlations. Short of the so-called *strong* minds-correlations, rejected by Hemmo and

Pitowsky themselves, the most obvious such assumption is of what we call *quasi-strong* correlations, which, however (or so we shall argue), are not justifiable within Albert and Loewer's theory. Finally, although we agree with point iii), we will qualify it slightly.

6.2 Albert and Loewer's Many Minds View.

6.2.1 Making sense of probabilities in an Everettian interpretation.

Albert and Loewer's purpose in elaborating the MMI was to present an interpretation of the Relative State Theory (Everett, 1957, 1973) which could represent a valid alternative to the view that worlds literally split put forward by DeWitt and Graham (DeWitt, 1971, and DeWitt and Graham, 1973). The latter view is not considered by Albert and Loewer to be an acceptable interpretation, due to three fundamental problems: the "democracy of basis problem", the "conservation of mass problem" and the "determinism problem".

The first problem consists in the fact that in the splitting worlds interpretation, the splitting of the original world seems to privilege a particular basis (the one corresponding to the measured observable), while the quantum mechanical formalism does not yield any privileged basis.

The "conservation of mass problem" concerns the fact that, according to Schrödinger's equation, the mass-energy of the combined observed system and measurement apparatus is the same before and after the measurement, while within the splitting worlds interpretation the total mass-energy increases after each splitting (after every measurement process).

The "determinism problem" calls into play the interpretation of probabilities. It is illustrated by Albert and Loewer in the following terms: "since, according to the [splitting worlds interpretation], it is certain that all outcomes of the measurement will occur and will be observed by successors of [the observer], what can be meant by saying that the probability of a particular outcome = c^2?" (1988, p. 201). Albert and Loewer's development of the MMI was driven by an attempt to solve these three problems and further the *mindless hulks problem*, which derives directly from the determinism problem and is the target of the subsequent discussion by Hemmo and Pitowsky.

According to Albert and Loewer it is impossible to make sense of probabilities in Everettian quantum mechanics by relying only on deterministic dynamical equations. A first problem is that the evolution of the wave function does not provide us with a rule for the transtemporal identity of branches. In DeWitt and Graham's interpretation, for instance, the only thing we can say about worlds is that, at a given instant *t*, to each component of the universal wave function corresponds a world in that state, but nothing enables us to say that the "worlds" existing at *t* are the same worlds that exist at a later instant *t'*. On the other hand, without transtemporal identity of branches there is no hope of making sense of statements like "the probability that *I* will register spin-up is = c^2", for nothing allows us to identify *me* before the measurement with any *me* existing after the measurement. Moreover, given the deterministic character of the dynamical equations, according to Albert and Loewer the addition of the transtemporal identity of minds is a necessary but not sufficient condition for a meaningful interpretation of probabilistic statements. This is the already mentioned "determinism problem": if all the outcomes of a measurement will occur with certainty, the probability for each outcome should be equal to 1, not to some c^2.

According to Albert and Loewer "if probability is to be introduced into the picture, it must necessarily be by *adding* something to the interpretation" (p. 201)—however, to "add something to the interpretation" typically means to forsake the central idea that the wave function is a complete description of the physical world—thus the problem arises of how to make sense of probability and, at the same time, maintain intact the Everettian postulate.

Keeping Albert and Loewer's view of the determinism problem in mind, we can now see their proposal for its solution. They begin with an intermediate construction, that of the "single mind" view.

6.2.2 The Single Mind View

The construction of this view starts with two basic postulates:

1. The universal wave function provides a complete physical description of reality.
2. Through introspection we are able to obtain reliable data regarding our beliefs.

Since introspection suggests that we always have well-defined beliefs, we infer that these cannot enter a superposed state.

According to postulate 1., our bodies are generally in a superposition of different brain states,[26] but according to postulate 2. our mind is never in a superposition of the corresponding belief states. This, in turn, implies that the following assumption fails:

[26] More generally, rather than of superpositions, one should always talk of "improper mixtures", since it is a larger system that enters the superposition, but we shall allow ourselves the slips in language.

M: “The state wherein A believes that spin-*x* = up and the state wherein A believes that spin-*x* = down are identical with certain physical states of A’s brain.”

That is, according to Albert and Loewer, *the desired theory has to be a dualist theory*.

The first proposal Albert and Loewer advance is the *single mind view*. Its basic postulates are:

a. The universal wave function provides a complete description of physical reality.
b. Every sentient physical system is associated with a nonphysical entity called *mind*, which is never in a superposition of belief states. Our state of consciousness corresponds to the state of our mind.
c. The evolution of the mind during measurements is genuinely stochastic. The probability for the mind to jump to a state after the measurement is given by Born’s rule, on the basis of the *local* (reduced) state of the observer.
d. Once a mind has jumped to a certain state, its successive evolution is ruled by the corresponding component of the state of the observer.

After each measurement the observer’s mind chooses only one of the component states of the observer’s brain, leaving the others “uninhabited”. (We shall elaborate on this point in our discussion of the MMI.)

Albert and Loewer do not characterise the kind of dualism to which the single mind theory is committed – all they say is that minds “are not quantum

mechanical systems; they are never in superpositions. This is what is meant by saying that they are non-physical" (Albert and Loewer, 1988, p. 207).[27]

Albert and Loewer acknowledge that, given the failure of M, a viable Everettian interpretation of quantum mechanics necessarily implies some violation of the supervenience of the mental on the physical. Nonetheless, this violation is considered too strong within the single mind theory.

The violation of supervenience emerges in two ways:

α) In order to say that minds evolve stochastically, we have to admit transtemporal identity for minds, but the latter is not determined by the physical evolution of the world. This violation of supervenience is considered unavoidable by Albert and Loewer, for it is essential to the solution of the determinism problem.[28]

β) Within the single mind view, the mental state does not supervene on the physical state since the same superposed physical state can correspond to different mental states.

While α) is considered by Albert and Loewer to be an unavoidable feature of a coherent Everettian theory, β) is regarded as too high a price to pay, and, in

[27] It is difficult to say whether the SMI implies property or substance dualism. Albert and Loewer seem to refer to a property dualism. However, it could be argued that if the single mind view wants to be explanatory of the apparent contradiction between the fact that our bodies are typically in superposition states and our minds are always in determined states of beliefs—then it forcedly turns into a *substance* dualism, for a weaker property dualism would not solve the tension between postulates 1. and 2.

[28] Albert and Loewer suggest another way, alternative to the many minds view, to restore it. This consists in adopting what they call the *Instantaneous Minds View*, which gets rid of the transtemporal identity of minds, i.e., in which there is no matter of fact regarding a unique relation of succession between the minds at an earlier and at a later time. If we renounce the transtemporal identity of minds, we have, at each instant, a set of minds which is completely determined by the brain state of the observer, and there is no reason of concern regarding the evolution of minds for the simple fact that there is nothing in the theory that allows us to talk about the evolution of each mind. However, Albert and Loewer reject also this possibility, because this would entail the impossibility of making sense of transition probabilities: "the cost of surrendering the 'trans-temporal identity of minds' would seem to be that we can no longer make sense of statements like 'the probability that *I* will observe spin up on measurement is *p*' since such statements seem to presuppose that it makes sense to talk of a single mind persisting through time" (Albert and Loewer, 1988, p. 211).

effect, an avoidable one: “on the single mind view, all but one of the elements of a superposition [of brain states] represent, as it were, mindless brains and which element represents a mind is not determined by the physical nature of the underlying brain state and cannot be deduced from the quantum state or from any physical experiment. The non-physicalism of the [single mind view] is especially pernicious. It entails that mental states do not even *supervene* on brain states (or physical states generally) since one cannot tell from the state of a brain what its single mind believes” (Albert and Loewer, 1988, p. 206).

The quotation above illustrates the *mindless hulks problem*, on the basis of which Albert and Loewer reject the single mind view. Now, given that Hemmo and Pitowsky’s argument on the weak nonlocality of the many minds view is based on Albert and Loewer’s position on the mindless hulks problem, we need to discuss this problem in more detail before proceeding to our criticism. Indeed, the many minds view was expressly proposed as a solution of the mindless hulks problem, and depending on what problem the theory is called to solve, it will be expected to have different features.

Note that often the mindless hulks problem is presented as follows. If only one component of Alice’s brain state is inhabited by her mind, and the same for Bob’s brain state, then Alice’s mind may be witnessing an uninhabited component of Bob’s physical state. In our opinion, however, and as suggested by the above quotation, Albert and Loewer’s worry about the mindless hulks problem is generated by the violation of supervenience entailed by the single mind view. Albert and Loewer feel compelled to bite the bullet of dualism and some non-supervenience of the mental on the physical, however they do not want to completely give up the idea of supervenience. The single mind view, in fact, does not only imply the failure of a *metaphysical* supervenience of the mental state on

the physical state maintained by physicalism, but even of a *nomological* supervenience, which is widely agreed upon even among dualist philosophers.

6.2.3 The Many Minds View

The many minds view is thus expressly proposed in order to make up for the "pernicious non-physicalism" displayed by the dualist character of the single mind view and exemplified by the mindless hulks problem. The single mind view's postulates are still valid, only in place of postulate b we have:

b'. Every sentient physical system is associated with *an infinity* of minds. If the observer's body is in a superposition state of beliefs, say $|B\rangle = c_1|B_1\rangle + c_2|B_2\rangle + ... + c_i|B_i\rangle$, the proportion of minds in state M_k (with 1<k<i), corresponding to B_k, is = $|c_k|^2$.

In a nutshell: an uncountable infinity of minds is tied to each brain state and at each measurement the set of minds splits in as many subsets as are the possible results of the measurement. The proportion of minds which ends up in a certain state after the measurement is equal to the squared coefficient of the corresponding brain state. The evolution of minds is still stochastic and governed by Born's rule, applied to the "inhabited" component of the state, thus allowing for the definition of conditional probabilities, say, in the case of successive measurements at times t_1 and t_2, for instance the conditional probability for a mind seeing up-up at t_2, given that it sees up at t_1.

Note that in order for later memories of observed results to faithfully track the inhabited components of the state, not only need the minds possess transtemporal identity, but one must also be able to reidentify the components of the physical state over time (the problem of the transtemporal identity of branches), which in

turn presupposes that the physical correlates of the mental states are subject to decoherence (Bacciagaluppi, 2003, section 4.3). This appears plausible, since they are the seat of memories, or at the very least are correlated with physical records in the environment.

We emphasise that the faithfulness of later reports is obviously relevant in the case of Bob's reports to Alice. Should Bob's report not be faithful in the sense above, then there is indeed a sense in which those of Alice's minds that witness Bob's report would be mistaken in attributing (past) mental states to Bob. And this could be taken as a variant of the mindless hulks problem. But if one takes decoherence explicitly into account, we believe this version of the problem is a red herring.

This brings us back to the mindless hulks problem and to our main line of enquiry,

namely: why should the many minds view be regarded as a solution to the mindless

hulks problem?

In the many minds view we can distinguish two different mental states: the *local mental state*, which is the state of individual minds, and the *global mental state*, i.e. the *distribution* of mental states among the infinity of the observer's minds (Albert and Loewer, 1988, and Barrett, 1999). Now, if, on the one hand, the local state does not supervene on the physical state (for individual minds evolve stochastically and we cannot deduce their state from the physical one), on the other hand the global mental state completely supervenes on the brain state, for the former is uniquely determined by the latter. The nomological supervenience of the mental on the physical is not completely restored; however, the problem is at least downgraded, for there is a sense in which the mental state

does supervene on the physical state. In this sense Albert and Loewer state: "we have purchased supervenience of the mental on the physical at the cost of postulating an infinity of minds associated with each sentient being" (Albert and Loewer, 1988, p. 207).

6.2.4 Locality: Albert and Loewer's argument.

In their 1988 article, where they first propose the many minds view, Albert and Loewer make a list of the merits of their theory, the last but not least of which is the fact that the many minds view provides an account in which all interactions are *local*. According to Albert, Bell's theorem has no significant consequences for the MMI, for "Bell proved that there can't be any local way of accounting for the observed correlations between the outcomes of measurements like that; but of course […] the idea that there ever *are* matters of fact about the "outcomes" of a pair of measurements like that is just what [the MMI] *denies*!" (Albert, 1992, p. 132).

Let us take a Bell type experiment. A pair of entangled particles, *a* and *b*, in state

$$\frac{1}{\sqrt{2}}\left(\left|+\right\rangle_a\left|-\right\rangle_b-\left|-\right\rangle_a\left|+\right\rangle_b\right) \qquad (1)$$

is sent to two distinct points C and D in space. Then Alice prepares to measure spin-x on *a* and Bob prepares to measure spin-(x+θ) on *b*, so that the state of the system composed by Alice, Bob, *a* and *b* at instant τ is:

$$\left|\Psi\right\rangle=\left|ready_x\right\rangle_A\left|ready_{x+\vartheta}\right\rangle_B\frac{1}{\sqrt{2}}\left(\left|+\right\rangle_a\left|-\right\rangle_b-\left|-\right\rangle_a\left|+\right\rangle_b\right) \qquad (2)$$

where, say, $\left|ready_x\right\rangle_A$ is the state of Alice who is ready to perform a measurement of spin-x. After the two measurements, at instant τ', Alice's and

Bob's bodies are entangled with *a* and *b*, thus the state of the composite system *a*+*b*+Alice+Bob becomes:

.

$$|\Psi\rangle' = \frac{1}{\sqrt{2}}\begin{pmatrix} \sin\frac{\theta}{2}|+_x\rangle_a|+_{x+\theta}\rangle_b|+_x\rangle_A|+_{x+\theta}\rangle_B + \cos\frac{\theta}{2}|+_x\rangle_a|-_{x+\theta}\rangle_b|+_x\rangle_A|-_{x+\theta}\rangle_B + \\ -\cos\frac{\theta}{2}|-_x\rangle_a|+_{x+\theta}\rangle_b|-_x\rangle_A|+_{x+\theta}\rangle_B - \sin\frac{\theta}{2}|-_x\rangle_a|-_{x+\theta}\rangle_b|-_x\rangle_A|-_{x+\theta}\rangle_B \end{pmatrix}$$

(3)

Here, the evolution of Alice's minds depends on the *local* (*reduced*) *state of Alice alone*, thus on the evolution of the state $|ready_x\rangle_A$ to the improper mixture

$$|A\rangle = \frac{1}{2}\left(|+_x\rangle_A\langle+_x|_A + |-_x\rangle_A\langle-_x|_A\right) \tag{4}$$

Therefore, one half of Alice's minds end up in the state "I am registering +", the other half end up in the state "I am registering -", *independently of Bob's measurement*. The same applies to Bob's minds. At this moment of the experiment there are thus no correlations between individual minds.

Let us now say that Alice and Bob meet and report their results to each other. At the new instant τ'' the state of the composite system has evolved to:

$$|\Psi\rangle'' = \frac{1}{\sqrt{2}}\begin{pmatrix} \sin\frac{\theta}{2}|+_x\rangle_a|+_{x+\theta}\rangle_b|+_x+'\rangle_A|+_{x+\theta}+'\rangle_B + \\ +\cos\frac{\theta}{2}|+_x\rangle_a|-_{x+\theta}\rangle_b|+_x-'\rangle_A|-_{x+\theta}+'\rangle_B + \\ -\cos\frac{\theta}{2}|-_x\rangle_a|+_{x+\theta}\rangle_b|-_x+'\rangle_A|+_{x+\theta}-'\rangle_B + \\ -\sin\frac{\theta}{2}|-_x\rangle_a|-_{x+\theta}\rangle_b|-_x-'\rangle_A|-_{x+\theta}-'\rangle_B \end{pmatrix} \tag{5}$$

where the state $|+_x+'\rangle_A$ is Alice's brain state that registers both spin-up on *a* and Bob's + report.

Here, after the interaction with Alice, $\cos^2\theta/2$ of Bob's minds believe that her result is the opposite of Bob's result. The probability for each of Bob's minds to

believe that the two results are opposite is $\cos^2\theta/2$, and the same is true of Alice's minds. Furthermore, each of Bob's minds that has registered $+_{x+\theta}$ has the conditional probability $\cos^2\theta/2$ of witnessing Alice's body reporting a - result. This is because the $|+_{x+\theta}>$ and $|-_{x+\theta}>$ components of Bob's state have decohered, and it makes sense to talk about the $|+_{x+\theta}>$ component of Bob's state splitting further into $|+_{x+\theta}+'>$ and $|+_{x+\theta}-'>$. Note that this is due to the *local* evolution of Bob's state when Bob meets Alice. Only at this point can we talk of correlations between subsets of minds (Albert and Loewer, 1988, p. 210, Bacciagaluppi, 2002, p. 111): Alice's set of minds in state $\left|+_{x}+'\right\rangle_{A}$ can be said to be "correlated" to Bob's set of minds in state $\left|+_{x+\theta}+'\right\rangle_{B}$. Note again that these "correlations" in the Bell type setting are reproduced by means of the local evolution of the brain states plus a completely local dynamics for the minds.

It is less misleading, however, to say simply that the two sets "correspond" to each other, in that the mental states in these sets encode appropriately matching results (in this case equal). The two sets have the same measure, so there exist ways of pairing off Alice minds with Bob minds in a measure-preserving way, but no particular way of doing so has any special significance. In particular, it is immaterial for the purposes of the MMI which of Bob's $+_{x+\theta}$ minds witness Alice reporting a + result and which witness Alice reporting a - result. Needless to say, we shall return to these "corresponding" sets below.

6.3 Hemmo and Pitowsky's argument.

In this section we shall finally present in detail Hemmo and Pitowsky's argument for the nonlocality of the MMI and our criticism of it.

We shall first clarify what Hemmo and Pitowsky mean by "weak minds-correlations". Then we shall discuss whether these correlations are nonlocal or not. This discussion will focus on the second point ii) of the three distinguished in the introduction: Hemmo and Pitowsky claim that from the sole assumptions of transtemporal identity of minds and of weak minds-correlations there follows a dependence of the partition of the sets of minds of an observer on the measurements performed by the other observers. Hemmo and Pitowsky present their supposed derivation of this dependence within their analysis of the MMI's description of the GHZ experiment. We will argue that weak minds-correlations are not sufficient for deriving this dependence, instead stronger assumptions are needed, for instance an assumption of quasi-strong minds-correlations, which holds counterfactually and is not deducible from the postulates of the MMI.

6.3.1 *Correlations*

Hemmo and Pitowsky's argument about the nonlocality of the MMI starts with this note about the mindless hulks problem: "it is not clear why [the mindless hulks] was considered to be a problem in the first place. The problem, recall, is that if each observer has only a single mind, then there is probability of one-half (given [an EPR-Bohm experiment with parallel settings]) that Alice's mind will follow the branch which is *not* followed by Bob's mind. For example, Alice's mind may perceive a – result and so may Bob's mind. When they meet, however, Alice's mind will with certainty perceive Bob as reporting a + result with certainty (this is the mindless hulk)" (Hemmo and Pitowsky, 2003, p. 237).

Briefly, in Hemmo and Pitowsky's analysis, Albert and Loewer's worry about the mindless hulks in the single mind view is given by the fact that every time I talk with a human being, there is the possibility that what I am interacting with is

not a sentient being, but a “mindless hulk”. In other words, in Hemmo and Pitowsky’s understanding, Albert and Loewer require of their theory that there is probability zero that the mind of a conscious being interacts with a mindless hulk. The lack of supervenience, which, in our view, is the core of the mindless hulks problem, is not considered here.

Now, Hemmo and Pitowsky say that there is a way to solve the mindless hulks problem within the single mind view, and that is to admit *strong correlations* between observers’ minds. Let us take the EPR experiment of section 3 and let us say that Alice and Bob measure the spin in the same direction. The strong correlations between minds guarantee that the results that Alice’s and Bob’s minds register are always opposite. In this way the mindless hulks problem in Hemmo and Pitowsky’s sense is solved. However, this kind of correlations would lead to a problem with nonlocality, for “Bell’s theorem implies that in a single mind theory the *correlations* between the minds of two observers on the two wings of a Bell-type experiment will satisfy the quantum predictions only by allowing strong nonlocal dependence between the trajectories of the minds” (Hemmo and Pitowsky, 2003, p. 228).

Here we just want to point out that this alternative solution to the mindless hulks is never considered by Albert and Loewer in their “step by step” construction of the MMI, and this for the good reason that *this alternative does not in fact solve the mindless hulks problem* as a problem of supervenience. Indeed, this alternative would make the problem of supervenience worse, since in the single mind view a correlation between minds could not be “mediated” or supported by the physical state, for minds do not supervene on it. But if the physical state is not able to support these correlations, a reference to a *direct*

connection between minds is necessary in order to account for them, and this would imply a definitely bizarre kind of dualism, which is hardly desirable.

Setting aside Albert and Loewer's intentions, can the above analysis be taken to justify the rejection of the single mind view? After all, the possibility that the mind of a conscious being interacts with a mindless hulk can be maintained coherently. Yet, so can solipsism. And the idea of a multitude of individual minds (i.e. a single mind for each individual) all leading independent existences seems to be a perverse kind of 'multiple solipsism'. Such an interpretation of the mindless hulks problem can thus indeed be used against the single mind view. We only wish to point out that this is not he same criticism as formulated by Albert and Loewer themselves. If interpreted in the sense of lack of supervenience, the mindless hulks problem must be considered completely solved within the MMI.

Back to the many minds view. According to Hemmo and Pitowsky, the many minds view does not exhibit the same strong correlations as the single minds view without mindless hulks problem would. However, they claim that Albert and Loewer's assumption of no correlations at all between observers' minds is incompatible with their position on the mindless hulks problem. Here we quote the passage where Hemmo and Pitowsky argue for the need for minds-correlations in the many minds view:

"According to Albert and Loewer there is no fact of the matter about whether or not Bob's + report, as perceived by Alice's − minds, corresponds to Bob's + minds. All we know is that Bob's report is associated with some minds, but Albert and Loewer do not allow us to say that these minds are + minds. Suppose there are only correlations between Alice's minds and Bob's reports (and vice versa), as Albert and Loewer say, but not between Alice's sets of minds and Bob's sets of minds. This means there will be a Bob + mind who witnesses Alice

reporting a − result, while the latter report is associated with a + mind of Alice. This is analogous to the mindless hulk problem. In the single mind case a mindless brain state is producing a definite − report, and in the many minds case a brain state associated with a + mind produces a − report. We believe that if the first is a problem, then so is the second.

In order to solve this problem we have to assume correlations between sets of minds as given by the quantum mechanical predictions. We call these correlations *weak minds-correlations*" (p. 237).

We have found this passage very obscure for a long time.[29] First of all, it is unclear whether Hemmo and Pitowsky think that the MMI needs to be modified by the introduction of weak minds-correlations; or whether they think that minds in the MMI are in fact weakly correlated, and that Albert and Loewer wrongly pretend that they are not (so that in fact Albert and Loewer wrongly claim that the MMI is local). Second, it is also unclear in what exactly consists Hemmo and Pitowsky's variant of the mindless hulks problem.

The worry about the mindless hulks is ostensibly that a Bob mind who thinks it is witnessing an Alice + mind is somehow instead "associated" with an Alice – mind. The most obvious way to read this worry is perhaps that, if one chooses any pairing off of Bob's minds with Alice's minds, if the stochastic evolution of the minds is *local*, then (whatever the probabilities for this evolution might be) the distribution of + minds and – minds cannot in general be equal to the quantum mechanical one, by Bell's theorem. Therefore what Bob's minds think Alice is reporting cannot in general be correct. The problem with this reading is that solving this variant of the problem would require the introduction of *strong* minds-correlations, as in Hemmo and Pitowsky's reading of the problem in the

[29] We are grateful to Meir Hemmo for correspondence on this point, which we think has clarified the issue at least in part.

single mind view. But this cannot be what Hemmo and Pitowsky have in mind for the MMI, since they themselves reject strong minds-correlations.

Incidentally, it seems out of place to consider *pairs* of Alice's and Bob's minds in the first instance. The picture given by MMI is that there are only two physical objects, Alice and Bob, which interact quantum mechanically and are accompanied each by an infinite set of minds. Each single Bob mind thus faces a single physical Alice body and an infinity of Alice minds, not any particular Alice mind. Simply, each of Bob's minds is aware only of a subset of Alice's minds, but attributes beliefs correctly to these minds (since they inhabit only the corresponding component of the state).

On the other hand, one could say, Bob's minds still face a "mindless hulk" if there is no direct connection between *any* of Alice's minds and Alice's report as witnessed by Bob's minds. Each of Bob's minds would have the *illusion* that it is interacting with an Alice mind behind the report it witnesses, and the proportion of Bob's minds that has a particular illusion matches the quantum mechanical prediction, as we have seen, so that the MMI recovers the appearences correctly. But if Alice's minds (of any kind) are just somehow "floating freely" in the background, then the recovery of the quantum mechanical predictions is indeed merely an illusion.

A more promising reading of Hemmo and Pitowsky's criticism is thus presumably to say that Bob's minds (or rather subsets thereof) should be thought of as interacting (in a way mediated by the relevant component of the quantum state) with precisely those of Alice's minds (or rather subsets thereof) of which they are aware. This would make any mismatch impossible, since Bob + minds witnessing Alice – reports would be interacting with and only with Alice – minds. Thus, instead of saying that Bob's minds are merely associated with Alice's

reports, as Albert and Loewer appear to be saying (at least according to Hemmo and Pitowsky), one could introduce "correlations" between the subset of Bob minds that witness a particular Alice report and the subset of Alice minds that have witnessed the corresponding result. Thus, on this reading, weak minds-correlations, which stave off a (perhaps rather distant) cousin of the mindless hulk problem, are available just below the surface of the MMI, but Albert and Loewer fail to capitalise on them.

The possibility of introducing weak minds-correlations, thus, is just a consequence of the dynamics of the minds described in section 2.3, including the conditional probabilities allowed by decoherence, i.e. it is the consequence of the evolution of each observer's set of minds, which is completely independent from the evolution of other observers' sets of minds.

Even on this reading, however, the form of the argument given by Hemmo and Pitowsky is still puzzling, because there is no obvious sense in which a Bob mind might get associated with an Alice mind of the "wrong" kind. If the worry is that a Bob mind may attribute beliefs to Alice that correspond to no Alice mind at all (in which case Bob's mind would indeed face a "mindless hulk"), then the worry is misplaced, since the form of the quantum state guarantees that there are Alice minds holding the belief attributed to Alice by the given Bob mind.

Apart from this slight remaining puzzle, on this reading Hemmo and Pitowsky's conclusion so far can be considered unquestionable, although instead of just saying that there are correlations between minds it might be more helpful to emphasise that there are correlations between physical reports, and that these correlations mediate those between the minds.[30]

[30] Indeed, it would be legitimate to ask how much a many minds interpretation seen from this perspective really differs from a (decoherence-based) many worlds interpretation augmented by one's favourite account of mental states. But this question would take us too far afield.

As a final point about the characterization of the weak minds-correlations, notice that if Alice and Bob meet and compare results, then the weak minds-correlations established at this point can be read backwards to the time when Alice and Bob originally perform their measurements: there will be subsets A_k of Alice minds and B_k of Bob minds which will "follow the same path" in the future. However (and crucially) the notion of weak minds-correlations allows us neither to specify in advance which subset B_k will accompany A_k in the future, nor to say that if Alice and Bob had performed different measurements than those actually performed, a given subset B'_k would have followed the same path as A_k. Given the stochastic dynamics of minds, in fact, *which* minds follow a determinate path is a contingent fact, and the same is valid for *which* minds will accompany A_k in the future and *which* minds would have accompanied A_k in a counterfactual situation.

6.3.2 Nonlocality.

So far, notwithstanding our reservations about details of Hemmo and Pitowsky's argument and about their motivation, we agree with their conclusion: within the many minds view, there are weak minds-correlations in the sense above. There remains the question of whether, as Hemmo and Pitowsky claim, these correlations are nonlocal.

According to Hemmo and Pitowsky, from the sole assumptions of transtemporal identity of minds and of weak minds-correlations there follows a dependence of the partition of the sets of minds of an observer on the measurements performed by the other observers. We will show how weak minds-correlations are not strong enough to deduce such a dependence of the partition of the minds, and how Hemmo and Pitowsky's argument can be saved by presupposing, instead, some quasi-strong minds-correlations.

First of all, let us consider again explicitly the Bell type situation of section 3. We have already argued there that the evolution of the minds is indeed completely local, while recovering at least the illusion of an interaction between the corresponding minds. The analysis above does not change if we consider explicitly also weak minds-correlations.

Indeed, it is true that the set of Alice's minds that (say) witness both a + result in Alice's experiment and Bob reporting a - result will in general depend on the setting $x + \theta$ chosen by Bob (indeed, the measure of this set will depend on it). However, this is the set of Alice's minds that *will* witness a - report when Alice meets up with Bob *in the future*, and there is no physical matter of fact about this set before that time, and so no nonlocal influence of Bob's choice on this partition of Alice's minds.

There is no influence of Bob's choice either on which of Alice's minds witnesses a + report in Alice's experiment, since, as long as later $\cos^2\theta/2$ of both Alice's + minds and Alice's - minds witness an opposite report from Bob, *which* of Alice's minds are going to witness + or - is immaterial (and similarly for Bob's minds). One could even imagine that the *same* half of Alice's minds always registers a + result independently not only of Bob's but even of Alice's own settings. Indeed, one might argue that the closest possible worlds to that with the actual settings are those in which the same minds witness + or - as in the actual world, even with a different setting.

Thus, in the standard EPR case weak minds-correlations do not require any form of nonlocality. Hemmo and Pitowsky's main argument, however, is phrased in terms of the GHZ experiment; so we turn to an explicit analysis of that experiment.

Consider the GHZ state:

$$\frac{1}{\sqrt{2}}\left(\left|+_z\right\rangle_1\left|+_z\right\rangle_2\left|+_z\right\rangle_3-\left|-_z\right\rangle_1\left|-_z\right\rangle_2\left|-_z\right\rangle_3\right) \qquad (6)$$

and let us say that three observers, Alice, Bob and Carol, respectively measure the spin of electrons 1, 2 and 3 in the same direction *x*. Because of the constraints dictated by quantum mechanics, the product of the spin of the three electrons in the direction *x* is always equal to -1. Always according to quantum mechanics, the state of the electrons and observers after the three measurements is:

$$\left|\Psi\right\rangle=\frac{1}{2}\begin{pmatrix}\left|+_x\right\rangle_1\left|-_x\right\rangle_2\left|+_x\right\rangle_3\left|+_x\right\rangle_A\left|-_x\right\rangle_B\left|+_x\right\rangle_C+\\ \left|-_x\right\rangle_1\left|+_x\right\rangle_2\left|+_x\right\rangle_3\left|-_x\right\rangle_A\left|+_x\right\rangle_B\left|+_x\right\rangle_C+\\ \left|+_x\right\rangle_1\left|+_x\right\rangle_2\left|-_x\right\rangle_3\left|+_x\right\rangle_A\left|+_x\right\rangle_B\left|-_x\right\rangle_C+\\ \left|-_x\right\rangle_1\left|-_x\right\rangle_2\left|-_x\right\rangle_3\left|-_x\right\rangle_A\left|-_x\right\rangle_B\left|-_x\right\rangle_C\end{pmatrix} \qquad (7)$$

Let us call this case *scenario 1,* and the cases in which Alice, Bob and Carol respectively measure the spin either in the directions (*x*, *y*, *y*) or (*y*, *x*, *y*) or (*y*, *y*, *x*), respectively, *scenarios 2, 3 and 4*. In these cases, the product of the results of the measurements should always be equal to 1.

According to Hemmo and Pitowsky, in the case of scenario 1, after the measurements the set M_a x M_b x M_c of *triples of minds* of Alice, Bob and Carol is partitioned into four subsets of triples that follow a specific component of the quantum state:

$$M_a\times M_b\times M_c=\begin{pmatrix}M^+{}_a(1)\times M^-{}_b(1)\times M^+{}_c(1)\cup\\ \cup\, M^-{}_a(1)\times M^+{}_b(1)\times M^+{}_c(1)\cup\\ \cup\, M^+{}_a(1)\times M^+{}_b(1)\times M^-{}_c(1)\cup\\ \cup\, M^-{}_a(1)\times M^-{}_b(1)\times M^-{}_c(1)\end{pmatrix} \qquad (8)$$

where $M^+{}_a(1)$ represents the set of Alice's minds which register + in scenario 1. The same is claimed to be true for every scenario (with the appropriate superposition state in the place of $\left|\Psi\right\rangle$ and, thus, the appropriate partition in place of (8)).

Given the above, Hemmo and Pitowsky now consider the sets:

$$M^{-}{}_{a}(1)\times M^{+}{}_{b}(1)\times M^{+}{}_{c}(1)\cap M^{-}{}_{a}(2)\times M^{+}{}_{b}(2)\times M^{-}{}_{c}(2)\cap \\ \cap M^{-}{}_{a}(3)\times M^{-}{}_{b}(3)\times M^{+}{}_{c}(3)\cap M^{+}{}_{a}(4)\times M^{+}{}_{b}(4)\times M^{+}{}_{c}(4) \qquad (9)$$

i.e. intersections of the sets of triples of minds that follow a specific component in each scenario.

There exist 4^4=256 possible combinations *of this kind*, in each of which there is an observer performing the *same* spin measurement in two different scenarios but obtaining *different* results (in the example of (9), it is Alice in scenarios 3 and 4). Now, according to Hemmo and Pitowsky, the 256 elements "exhaust every logical possibility. Therefore, at least one of those sets has probability ≥ 1/256" (p. 241).[31]

If this is correct, then it is necessarily the case that a non-zero measure set of minds of at least one observer registers *different measurement results* (for the same measurement), depending on the *settings* chosen by the other observers, which would indeed be a form of nonlocality (called weak nonlocality by Hemmo and Pitowsky).

What are we to make of this argument? First of all we note an inaccuracy in Hemmo and Pitowsky's presentation. They define $M^{+}{}_{a}(1)$ as the set of Alice's + minds after her measurement of *x*-spin in scenario 1 (and the other sets accordingly). If, however, a set of the form $M^{+}{}_{a}(1)$ x $M^{-}{}_{b}(1)$ x $M^{+}{}_{c}(1)$ is meant to represent minds that have followed the same component of the state, then $M^{+}{}_{a}(1)$ cannot denote the same set in $M^{+}{}_{a}(1)$ x $M^{-}{}_{b}(1)$ x $M^{+}{}_{c}(1)$ and in $M^{+}{}_{a}(1)$ x $M^{+}{}_{b}(1)$ x $M^{-}{}_{c}(1)$.

[31] At first sight, one might wonder about which probability measure is being used here, given that these sets are defined with reference to different scenarios (in which different quantum states apply). But a set of the form (11) is just a subset of M_a x M_b x M_c, and any measure on M_a x M_b x M_c for which such sets are measurable will do.

Hemmo and Pitowsky must mean in the first case the set of Alice's + minds in scenario 1 that will witness Bob and Carol reporting - and +, respectively (if Alice, Bob and Carol subsequently compare results), and in the second case the set of Alice's + minds in scenario 1 that will witness Bob and Carol reporting + and -. Equation (8) should thus be rewritten as:

$$M_a \times M_b \times M_c = \begin{pmatrix} M^{+-+}{}_a(1) \times M^{+-+}{}_b(1) \times M^{+-+}{}_c(1) \cup \\ \cup M^{-++}{}_a(1) \times M^{-++}{}_b(1) \times M^{-++}{}_c(1) \cup \\ \cup M^{++-}{}_a(1) \times M^{++-}{}_b(1) \times M^{++-}{}_c(1) \cup \\ \cup M^{---}{}_a(1) \times M^{---}{}_b(1) \times M^{---}{}_c(1) \end{pmatrix} \qquad (10)$$

where $M^{+-+}{}_a(1) \cup M^{++-}{}_a(1) \cup M^{-++}{}_a(1) \cup M^{---}{}_a(1)$ is in fact a partition of Alice's minds in scenario 1 (the other sets $M^{+++}{}_a(1)$, etc., are all empty, while the sets in the partition each contain exactly ¼ of Alice's minds). And similarly for Bob's and Carol's minds.

This, however, does not help with the next, crucial problem with (8). Namely, whether we consider the sets in (8) or in (10), they *do not form a partition* of M_a x M_b x M_c (which, remember, is defined as the set of the *triples of minds*). Indeed, other subsets of M_a x M_b x M_c such as $M^{+-+}{}_a(1)$ x $M^{-++}{}_b(1)$ x $M^{++-}{}_c(1)$ are not empty, simply because none of the factors $M^{+-+}{}_a(1)$, etc., are.

Again, Hemmo and Pitowsky must mean something else: namely, that a set such as $M^{+-+}{}_a(1)$ x $M^{-++}{}_b(1)$ x $M^{++-}{}_c(1)$, which does not correspond to any single component of the state, i.e. a set collecting minds together in triples whose records do not match, is given measure zero by some appropriate measure. One could certainly have this if one considered strong minds-correlations, i.e. correlations at the level of the stochastic evolution of the individual minds. In this case, any triple of minds could indeed be taken to have probability zero of ending

up with records that do not match. But strong minds-correlations are precisely the kind of correlations also Hemmo and Pitowsky wish to avoid. And the assumption of weak minds-correlations (as we understand them) merely requires that there be sets of minds such as $M^{+-+}{}_a(1)$, $M^{+-+}{}_b(1)$ and $M^{+-+}{}_c(1)$, which do match in their records and have the same size (in this case ¼).

Thus, Hemmo and Pitowsky's argument proceeding from consideration of the sets (9) has a false premise: that the sets defined as in (8) form a partition of M_a x M_b x M_c. Therefore the 256 sets of the form (9) do not form a partition of M_a x M_b x M_c either, and the argument to the effect that at least one of them has non-zero measure fails.

Instead, the requirement of weak minds-correlations is compatible with all of the sets of the form (9) being empty. Take the set $M^{-}{}_a(1) \times M^{+}{}_b(1) \times M^{+}{}_c(1)$, or more precisely the set $M^{-++}{}_a(1)$ x $M^{-++}{}_b(1)$ x $M^{-++}{}_c(1)$. As we have seen in the previous section, the weak minds-correlations imply that, whatever results the minds in $M^{-++}{}_a(1)$ will register in scenarios 2, 3 and 4, there will always be *some* minds of Bob and *some* minds of Carol that "keep them company". However, the weak minds-correlations do not require any of those minds to be elements of $M^{-++}{}_b(1)$ or of $M^{-++}{}_c(1)$. If indeed they are not, all the 256 intersections like (9) are empty, and no mind flips its sign depending on the measurement performed by other observers. And as we have pointed out in the case of the EPR scenario, one can maintain that the same minds register + (or -) in all scenarios, so that no nonlocality arises at the stage of the measurements. The only requirement imposed by weak minds-correlations is that, say, in scenario 1 Alice's + minds then divide evenly into +-+ minds and ++- minds, etc. This further evolution, however, takes place only when Alice meets Bob and/or Carol (or otherwise learns about their results) and is thus also to be thought of as local.

Thus, the claim that at least some of the 256 combinations are non-empty (and have non-zero measure) cannot be grounded in the weak minds-correlations, but only on some stronger assumption. If we discard strong minds-correlations, the easiest way of ensuring this appears to be the assumption that *the same minds be weakly correlated in all scenarios.* Thus, for instance, the intersection (9) would turn out to be just the set $M^{-}_{a}(1) \times M^{+}_{b}(1) \times M^{+}_{c}(1)$, that is, more accurately notated, the set $M^{-++}_{a}(1) \times M^{-++}_{b}(1) \times M^{-++}_{c}(1)$. We shall call this an assumption of *quasi-strong minds-correlations*. Note that it is entirely formulated at the level of counterfactuals. Note also that it does not force the sets in (9) to form a partition of $M_a \times M_b \times M_c$, nor *a fortiori* the sets in (8). In this sense it is a weaker assumption than that of strong minds-correlations.

It is not clear whether such an assumption would be acceptable to Hemmo and Pitowsky, and if so whether they would consider it justifiable within the MMI. It is certainly compatible with the postulates of the MMI, and could be added to it, perhaps in the more general (and to be made precise) form that as many minds as possible that are weakly correlated in the actual world should be weakly correlated in counterfactual situations. This could be read as a requirement on the closeness of possible worlds (but so is the requirement that the results witnessed by individual minds should not flip). Or its violation could be read as a distant cousin of the mindless hulks problem as understood by Hemmo and Pitowsky: Alice's minds will interact with different sets of minds depending on which scenario is chosen for the measurements. But we doubt this would cause Alice's minds very great existential pangs. Thus, we conclude that neither strong minds-correlations nor quasi-strong minds-correlations are required in Albert and Loewer's MMI, so that Hemmo and Pitowsky's accusation of nonlocality can be rejected.

As a final remark (almost a footnote): even on the assumption of quasi-strong minds-correlations, one might wonder how to characterise the dependence of the partition of the minds on the settings chosen by the other observers, i.e. whether it is indeed nonlocal (Hemmo and Pitowsky's point iii)). Indeed, one should take into account also the cases in which none of the observers meet or otherwise learn of one another's results. In this case, one might argue that the requirement of quasi-strong minds-correlations is vacuous. At the very least, one might argue that the constraint comes in place only at a later time, when the observers learn of one another's results, so that the most natural interpretation of the constraint may be not in terms of action-at-a-distance (spacelike) but of backwards causation (timelike), thus arguably more compatible with relativity. Since the issue is controversial, however, we shall not press this point.

Appendix 1

The Einstein-Podolsky-Rosen Argument

The Einstein-Podolsky-Rosen (EPR) argument was formulated in 1935 in the paper *Can quantum-mechanical description of physical reality be considered complete?* as a proof against the completeness of quantum mechanics. It has been of crucial importance in the further development of the research on quantum mechanics, although in the immediate aftermath its publication it has been underestimated by many great physicians.

The argument formulated by EPR is reductio ad absurdum starting from the assumption that quantum mechanics is complete, i.e. that every element of reality has its representative in the theory, reaching then a contradiction and demonstrating in this way that the assumption is wrong.

The starting point is what has been called the EPR *Criterion of Reality*:

If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of reality corresponding to that quantity.

For pedagogical purposes, we will use the more simple version of the argument provided by Bohm (1951). Let's suppose that a particle whose total spin angular momentum is zero decays into two particles *a* and *b*. Due to the law of conservation of energy, the result of a measurement of *a*'s spin in a generic direction x is up iff the measure of the spin on *b* in the same direction yields down. For the assumption of the completeness of quantum mechanics we can say that the state of the composed system *a*+*b* is completely described by the non-factorizable state vector:

$$|\phi\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow_n\rangle_a |\downarrow_n\rangle_b - |\downarrow_n\rangle_a |\uparrow_n\rangle_b\right) \qquad (1)$$

where *a* and *b*'s states are entangled.

Let's further suppose that after the emission, the two particles are completely isolated one from the other, in such a way that they cannot interfere. In order to do that it will be sufficient to let them move in opposed directions and that they reach an adequate distance. At this point two observers, Alice and Bob, measure the spin respectively of *a* and *b* in the same direction n. At the instant τ_1 Alice performs the measure on *a*: if the result is spin-n=up, then the state of *a*+*b* instantaneously reduces to:

$$|\phi\rangle' = |\uparrow_n\rangle_a |\downarrow_n\rangle_b \qquad (2)$$

At this point Alice knows with certainty (i.e. with probability equal to 1) that Bob's measure on *b* will be spin-n=down. If the result obtained by Alice is spin-n=down, then the state of *a*+*b* reduces to the second term of (1) and the result of Bob's measurement will be spin-n=up with certainty.

The independence of *b*‘s state from Alice’s knowledge state is a consequence of the Principle of Locality, stating that two “objects (A and B) far apart in space: external influence on A has no direct influence on B.” (Born, 1971, pp. 170-71)

Figure 1 represents the light cones having as origins respectively the events “Alice performs a measurement of spin-n on *a*” (S) and “*b* have a definite spin-n value” (B). Each event lies outside the light cone of the other, which means that the spatio-temporal interval between the two events is a space-like interval , from which it follows that (S) and (B) cannot be causally related. Given the Criterion of Reality, *b*‘s spin-n value must be an objective poperty of *b.*

On the other hand, if this is so, *b*’s spin-n property should have been objective also before τ_1, since no disturbance could occur on *b* to modify its state. as a consequence, *a* and *b* should be in an eingenstate of the observable spin-n already some instant before τ_1, which, however, contradicts the initial assumption describing *a*+*b*’s state as non-factorizable.

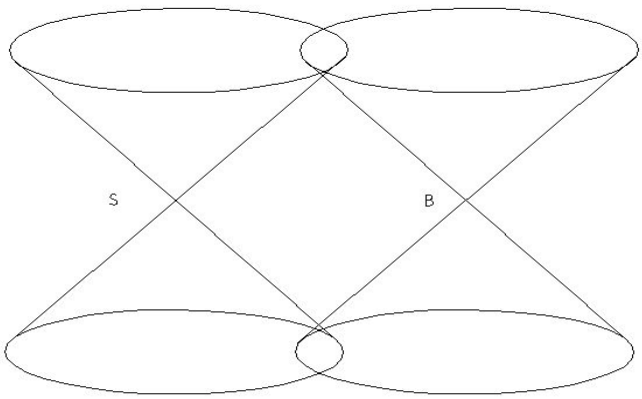

figure 1

Finally, the EPR argue that at τ_1 Alice could have decided to measure *a*'s spin in any other direction, and in any case she could have predict the value of the chosen observable in *b*; but if this is true, then *b* must posses a determinate value for every observables.

As a conclusion, the EPR argument is grounded on three assumptions:

- the validity of the Criterion of reality;

- the validity of the Principle of Locality;

- the validity of the principle of completeness;

according to a more modern analysis of the argument, the contradiction to which the three assumptions demonstrate that they are not compatible and therefore that quantum mechanics must violate at least one of these three. EPR though that the contradiction should have been solved by eliminating the assumption of the validity of the principle of completeness, however, in 1964 John Bell demonstrated that what is in reality violated by quantum mechanics is the locality principle.

Appendix 2

Bell's Theorem

Bell's demonstration of the non local feature of quantum mechanics starts from his mathematical formulation of the principle of locality.

Let's take two zones A and B of space-time, separated by a space-like interval (figure 2).

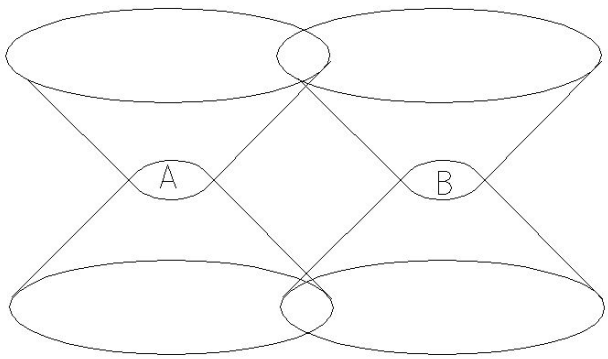

Figure 2

According to SR, the events lying in A cannot directly influence those in B, and vice versa.

This does not mean that there cannot be a correlation between the events belonging to the two cones. A and B's past cones overlap in the region C, where can be events that influenced the occurrence of both A and B. As a consequence, if the available information on A's absolute past are not complete, we can add to them available information on B's events. The latter, in fact inform on their causes, and between them there can possibly be events belonging to C.

Let's see a specific case: let o be the set of all the events belonging to C; let α be the set of some events of A's absolute past cone (C excluded) and let β be the set of events belonging to B. From the events belonging to β which have some common causes with some event in A (i.e., those inside C) we can infer information on A. On the other hand, these information already present in o. As a consequence:

$$prob(A:\alpha,o) = prob(A:\alpha,o,\beta) \tag{3}$$

Where the left side of the equation expressses the probability assigned by α and o events to A events. A theory which satisfies (3) is said by Bell to be locally causal. From the above definition it can be demonstrate that quantum mechanics in his standard version is non local.

Let *a* be a particle in state

$$\frac{1}{\sqrt{2}}|c\rangle_a + \frac{1}{\sqrt{2}}|d\rangle_a \tag{4}$$

At the instant τ, where $|c\rangle_a$ is *a*'s state describing the event c_a: "*a* is in the point *c* of space", while $|d\rangle_a$ is the state describing the event d_a: "*a* is in point *d* in space". Regions C e D of space-time, where c_a and d_a are, are separated by a space-like interval. Let's furthermore call E the events belonging to the overlap of the two past cones and K the events of the rest of C past cone and d some events of region D (figure 3).

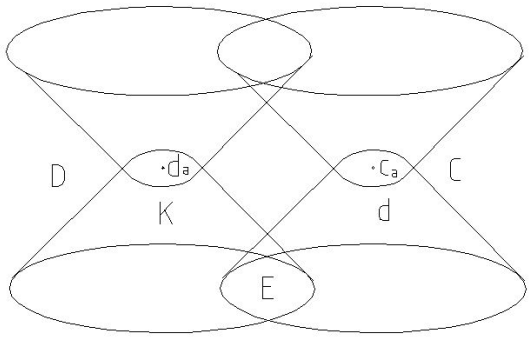

Figure 3.

Due to the principle of locality and given the assumption of the completeness of quantum mechanics, then:

$$prob(c_a; k, E) = prob(c_a; k, E, d) \,, \tag{5}$$

i.e., given the events belonging to K and E, the probability of c_a is equal to the probabilità of c_a given the events in its past light cone plus those in a region separated from a space-like interval.

On the other hand, as it is known this is not what quantum mechanics predicts, and in fact, if we verify if *a* is in *d*, and we effectively find it there, then *a*'s state instantaneously collapses in $|d\rangle_a$ and the c_a's probability becomes equal to zero. If, on the other hand, no measures are performed, then *a*'s state remains (4) and c_a's probability is ½.

Before concluding that the locality principle does not hold in quantum mechanics, there is still an alternative to consider. There is still the possibility that quantum mechanical probabilities depends on the incompleteness of the quantum state, which, therefore, would not determine all the real properties of a system. In such a case the locality principle could still hold.

Bell's theorem is the result of the attempt to verify the efficacy of such a hypothesis. It demonstrates that, if quantum mechanical predictions are exact, not only quantum mechanics is not local, but no interpretation can make it local.

Take two particles a and b in entangled state:

$$|\varphi\rangle = \frac{1}{\sqrt{2}}\left(|\uparrow_x\rangle_a |\downarrow_x\rangle_b - |\downarrow_x\rangle_a |\uparrow_x\rangle_b\right) \qquad (6)$$

Moving in opposite directions. Two measurement apparatus measure the spin of the two particles in the space-time zone A and B. One can measure the spin in direction *u* or *p*, the other in direction *r* or *t*.

Bell examines a similar experiment, repeated a sufficient number of times, and wonders what kind of correlations can have the results of the measures at the two edges of the experiment, given that the principle of locality holds. In other words, Bell wonders what kind of correlations should have the probabilities of the results of each measurement in a theory where: 1) the results of the measure on a particle cannot influence those on the other particle; 2) the state of the electrons decides all the objective properties of the particles. From these two conditions alone Bell derive his famous inequalities, whose arguments are the prediction of the long term average results of the measurements.

Finally, Bell demonstrate that quantum mechanical predictions do not respect the obtained inequalities.

Notice that: 1) the premises of the theorem do not lie on any specific interpretation of quantum mechanics, but refer to any theory that reproduces quantum mechanical average values; 2) the above predictions have been widely verified and it can be said that no doubt can be advanced against their validity. This in turn implies that the example just illustrated cannot be described by a theory which respect the locality principle.

Bibliography.

Achinstein, P., (1983) *The Nature of Explanation*, Oxford: Oxford University Press.

Albert, D., 1992, *Quantum Mechanics and Experience*, Cambridge-London, Harvard University Press.

Albert, D., and Loewer, B., 1988, "Interpreting the Many Worlds Interpretations", in *Synthese*, LXXXII, n. 2, pp. 195-213.

Bacciagaluppi, G. (1994). Separation theorems and Bell inequalities in algebraic quantum mechanics. In P. Busch, P. Lahti, and P. Mittelstaedt (eds.), *Symposium on the foundations of modern physics 1993: quantum measurement, irreversibility and the physics of information*. pp. 29–37. Singapore: World Scientific.

Bacciagaluppi, G., 2002, "Remarks on Space-time and Locality in Everett's Interpretation", in *Non-Locality and Modality*, eds. T. Placek and J. Butterfield, Dordrecht, Kluwer. Also in *PhilSci Archive*, http://philsci-archive.pitt.edu/archive/00000504/01/cracow.pdf.

Bacciagaluppi, G., 2003, "The Role of Decoherence in Quantum Mechanics", *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/qm-decoherence/.

Baker, A. (2005). “Are there genuine mathematical explanations of physical phenomena?” *Mind*, 114(454), 223-238

Baez J. (2006), “Quantum Quandaries: a Categorical-Theoretic Perspective”, in S. French, D. Rickles, J. Saatsi (eds.), *Structural Foundations of Quantum Gravity*, Oxford University Press, pp. 240-265.

Barnum, H., Caves, C. M., Fuchs, C. A., Jozsa, R., and Schumacher, B. (1996). Noncommuting mixed states cannot be broadcast, *Phys. Rev. Lett.*, 76:2318.

Berger, R. (1998) “Understanding Science: Why Causes Are Not Enough”, *Philosophy of Science*, 65, pp. 306-332

Barrett, J. A., 1999, *The Quantum Mechanics of Minds and Worlds*, Oxford, Oxford University Press.

Bohm, D., 1951, *Quantum Theory*, New York: Prentice Hall

Boltzmann, Ludwig. 1964. *Lectures on Gas Theory*. Berkeley: University of California Press.

Born, M., (ed.), 1971, *The Born-Einstein Letters*, New York; Walker.

Bradie, M., (1996), “Ontic Realism and Scientific Explanation”, *Philosophy of Science*, vol. 63, supplement. Proceedings of the 1996 Biennal Meeting of the Philosophy of Science Association. Part I. Contributed Papers (Sep., 1996), pp. S315-321.

Brandom R. (1998), *Making it Explicit*, Harvard University Press, Harvard.

Bricmont, (1995). “Science of Chaos or Chaos in Science?” *Physicalia Magazine*, 17, (1995) 3-4, pp.159-208. http://dogma.free.fr/txt/JB-Chaos.htm

Brown, H. (2005), *Physical Relativity. Space-time Structure from a Dynamical Perspective*, Oxford University Press

Brown, H. and Timpson, C. (2006), ‘Why Special Relativity Should Not Be a Template for a Fundamental Reformulation of Quantum Mechanics’, in *Physical Theory and its Interpretation: Essays in Honor of Jeffrey Bub*, Demopoulos and Pitowsky (eds.) Springer, http://xxx.lanl.gov/abs/quant-ph/0601182.

___,Pooley, O., (2006), "Minkowski Space-Time: a glorious non-entity", in Dieks D. (ed.), *The Ontology of Spacetime*, Elsevier, pp. 67-89.

Bub, J., (1974), The Interpretation of Quantum Mechanics, Dordrecht, Holland: Reidel.

----, (1979) "The Measurement Problem of Quantum Mechanics", Problems in the Philosophy of Physics, Bologna: Società Italiana di Fisica, pp. 100-104.

Bub, J. (1997), Interpreting the Quantum World, Cambridge: Cambridge University Press.

Bub, J. (2000), 'Quantum Mechanics As A Principle Theory', *Studies in the History and Philosophy of Modern Physics*, 31B, 75–94.

Bub, J. (2004), 'Why the Quantum?' *Studies in the History and Philosophy of Modern Physics*, 35B, 241–266. arXiv:quant-ph/0402149 v1

Bub, J. (2005), 'Quantum Theory Is About Quantum Information', *Foundations of Physics*, 35(4), 541–560. arXiv:quant-ph/0408020 v2

Cartwright, Nancy (1983), *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Clifton, R., 1998, "Structural Explanation in Quantum Theory" http://philsci-archive.pitt.edu/archive/00000091/00/explanation-in-QT.pdf

Clifton, R., Bub, J. and , Halvorson, H. (2003), 'Characterizing Quantum Theory in Terms of Information–Theoretic Constraints', *Foundations of Physics*, 33(11), 1561–1591.

Contessa, G., (2007), "Scientific Representation, Interpretation, and Surrogative Reasoning", *Philosophy of Science*, 74 (January 2007) pp. 48-68. http://philsci-archive.pitt.edu/archive/00003291/01/Paper_-_Representation,_Interpretation,_and_Surrogative_Reasoning_(PhilSci).pdf

de Regt, H.W. (2004), "Discussion Note: Making Sense of Understanding", *Philosophy of Science* 71 (1): 98–109.

de Regt, H.W., Dieks, D., (2005), "A Contextual Approach to Scientific Understanding", *Synthése*, 144, n.1., pp. 137-170.

DeWitt, B. S., 1971, "The Many-Universes Interpretation of Quantum Mechanics", in *Foundations of Quantum Mechanics*, New York, Academic Press. Reprinted in DeWitt and Graham (1973), pp. 167-218.

DeWitt, B. S., and Graham, N. (eds.), 1973, *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton, Princeton University Press.

Di Salle R. (1995), "Spacetime theory as physical geometry", *Erkenntnis*, 42, pp.317-337.

Duhem, P., (1954) *The Aim and Structure of Physical Theory*, Princeton : Princeton University Press.

Einstein, A. (1954) "What is the theory of relativity." *Ideas and Opinions*, NewYork: Bonanza Books, pp. 227–232. First published in *The Times*, London, November 28, 1919, p. 13.

___, Podolsky, B. and Rosen, N., (1935) Can quantum-mechanical description of physical reality be considered complete? Phys. Rev. 47 777

Everett, H. III, 1957, "'Relative State' Formulation of Quantum Mechanics", in *Reviews of Modern Physics*, XXIX, pp. 454-462.

Everett, H. III, 1973, "The Theory of the Universal Wave Function", in DeWitt and Graham (1973), pp. 1-140.

Feynman R. (1965), *The Character of Physical Law*, Cambridge Mass, M.I.T. Press

Fine A. (1989), "Do Correlations need to be explained'", in Cushing J. McMullin E. (eds.) *Philosophical Consequences of Quantum Theory*, Notre Dame University Press, Notre Dame, IN, pp.175-194

French, S., (1999) 'Models and Mathematics in Physics: The Role of Group Theory', in *From Physics to Philosophy*, ed. J. Butterfield and C. Pagonis (Cambridge: Cambridge University Press), pp.187–207.

Friedman, M., 1974, 'Explanation and Scientific Understanding', *Journal of Philosophy*, 71: 5-1

Frigg, R., (2006), "Scientific Representation and the Semantic View of Theories", *Theoria* 55: 37-53. http://philsci-archive.pitt.edu/archive/00002926/01/Scientific_Representation.pdf.

Goodman, N. (1968), *Languages of Art.* Indianapolis: Bobbs Merrill.

Halvorson, H., (2003). 'A note on information-theoretic characterizations of physical theories', Studies in History and Philosophy of Modern Physics 35, 277-293, quant-ph/0310101.

Hartmann, Stephan. 1999. "Models and stories in hadron physics." Pp. 326-346 in *Models as Mediators*. Edited by M. Morrison and M. Morgan. Cambridge: Cambridge University Press. http://philsci-archive.pitt.edu/archive/00002433/01/Stories.pdf

Heisenberg, W. (1927) 'Ueber den anschaulichen Inhalt der quantentheoretischen Kinematik and Mechanik' *Zeitschrift für Physik* **43** 172-198. English translation in Wheeler and Zurek, 1983, W.H. (eds) (1983) *Quantum Theory and Measurement* (Princeton NJ: Princeton University Press), pp. 62-84.

Hemmo, M., and Pitowsky, I., 2003, "Probability and Nonlocality in Many Minds Interpretations of Quantum Mechanics", in *British Journal for the Philosophy of Science*, 54, n. 2, pp. 225-243.

Hemmo, M. and Hagar, A., "Foundations of Physics, Volume 36, Number 9, September 2006 , pp. 1295-1324(30)

Hempel, C. G, and Oppenheim, P., (1948), “Studies in the Logic of Explanation”, Philosophy of Science 15, pp. 135-175. Reprinted in (Hempel, 1965).

Hempel, C., (1965), Aspects of Scientific Explanation and Other Essays in the Philosophy of Science, New York, The Free Press.

Hilgevoord, J. and Uffink J., "The Uncertainty Principle", *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2006/entries/qt-uncertainty/>.

Hitchcock, C., 1995, ‘Discussion: Salmon on Explanatory Relevance.’, *Philosophy of Science* 62: 304-20.

Hughes, R.I.G. (1989a), “Bell’s theorem, ideology, and structural explanation” , in Cushing, J. and McMullin, J. eds, *Philosophical Consequences of Quantum Theory.* Notre Dame, Ind. 1989, pp.

----, 1989b, *The Structure and Interpretation of Quantum Mechanics*, Harvard University Press.

----(1993), *Theoretical Explanation*, Midwest Studies in Philosophy XVIII: 132-153.

Hughes, R.I.G., (1997) *Models and Representation*, Philosophy of Science, vol. 64, No. 4, Supplement. Proceedings of the 1996 Biennial Meetings of the philosophy of Science Association. Part II: Symposia Papers, pp.325-336

Kac, M., (1959) *Probability and Related Topics in the Physical Sciences*, Interscience Pub., New York.

Kitcher, P., 1989, ‘Explanatory Unification and the Causal Structure of the World’, in (Kitcher and Salmon, 1989)

Kitcher, P. and Salmon, W., (1987), “van Fraassen on Explanation”, Journal of Philosophy, 84, pp.315-330.

___(eds.) (1989), *Scientific Explanation,* 410-505. Minneapolis: University of Minnesota Press.

Körner, S., (1969). *Fundamental questions of philosophy*, Harmondsworth, Middx.: Penguin Books.

Landau, L.J. (1987). On the violation of Bell's inequality in quantum theory, *Physics Letters A, 120*, 54–56.

McCulloch, W. S. and W. H. Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* **7**, pp. 115-133.

Piccinini, G., (2006), "Computational explanation in neuroscience", *Synthése*, 153, N. 3

Reichenbach, H., (1956). *The Direction of Time*. University of California Press.

Resnik M. (1981), "Mathematics as a Science of Patterns: Ontology and Reference", *Nous*, vol.15, 529-550.

Robertson J.P, (1929), "The Uncertainty Principle", *Physical Review*, 34, pp. 163–164.

Salmon, W., 1998, *Causality and Explanation*, Oxford: Oxford University Press.

___, 1989, *Four Decades of Scientific Explanation*, Minneapolis:University of Minnesota Press.

___(1994), "Causality without Counterfactual", Philosophy of Science 61:279-312

___(1984), "Scientific Explanation and the Causal Structure of the World", Princeton, Princeton University Press.

Schrödinger, E. (1936). Probability relations between separated systems. Proc. Camb. Phil. Soc., 32:446–452.

Schurz, G., and Lambert, K. (1994), "Outline of a Theory of Scientific Understanding", Synthese 101: 65–120.

Shapiro, S. (1983), "Mathematics and Reality", *Philosophy of Science*, vol.50, 523-548

___(1997), *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press.

Shapiro, S (2000), Thinking about mathematics, Oxford University Press.

Steiner, M., 1978, “Mathematical Explanation”, *Philosophical Studies*, (Minneapolis), 34:2, pp.135-151.

Suárez, M., (2004) An Inferential Conception of Scientific Representation, Philosophy of Science, 71 pp. 767–779. http://philsci-archive.pitt.edu/archive/00000991/

Summers, S. (1990). On the independence of local algebras in quantum field theory, *Reviews in Mathematical Physics, 2*, 201–247.

Swoyer, Chris (1991), "Structural Representation and Surrogative Reasoning," *Synthese* 87: 449–508.

Timpson, C. G., (2004). Quantum information theory and the Foundations of Quantum Mechanics. http://philsci-archive.pitt.edu/archive/00002344/01/qinfandfoundsqm.pdf

Trout, J.D., (2002), “Scientific Explanation And The Sense Of Understanding” Philosophy of Science***, 69 pp. 212–233***

___(2005), “Paying the Price for a Theory of Explanation: De Regt’s Discussion of Trout” *Philosophy of Science*, 72 pp. 198–20

van Fraassen, Bas C. (1980), *The Scientific Image*. Oxford: Oxford University Press.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA, MIT Press.

Woodward, James (2003), *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

_____(2003), "Scientific Explanation", *The Stanford Encyclopedia of Philosophy (Summer 2003 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2003/entries/scientific-explanation/>.