EMERGENCE IN THE PHILOSOPHY OF MIND

Markus Eronen
Department of Philosophy
University of Helsinki
Master's Thesis
November 2004

CONTENTS

1. Introduction

The concept of emergence has found its way back to the mainstream of philosophy. The air of mysticism that earlier surrounded the concept has disappeared, and it is no longer considered dubious to use expressions like “emergent properties” or “emergent phenomena”.

The tradition of British Emergentism that began with John Stuart Mill faded before the middle of the 20th century when positivist and reductionist ideas started to dominate the field of philosophy. However, by the 1970s it was becoming clear that the reductionist approaches could not convincingly account for mental phenomena. This lead to the development of different nonreductive theories and the return of emergentism.

The most central concept in this new emergentism is *irreducibility*. The idea is that although mental properties depend on physical properties and supervene on them, they can never be reduced to them. This idea is also evident in the works of the British Emergentists, particularly in C. D. Broad’s *The Mind and its Place in Nature* (1925). The high-flown evolutionary and cosmological theories of the classic emergentists that are probably the reason for the bad reputation of emergentism are not a part of the current debate.

The most difficult problem that a current emergentist has to face is the problem of mental causation. The problem is this: if the world is fundamentally physical, as emergentism supposes, how can emergent mental properties have causal powers? If they have a role in causing physical events, it seems that physical events have causes that are outside the scope of physics, and physics alone is not enough to explain all physical events. This is an unacceptable outcome. If emergent mental properties don’t have causal powers, it is not clear in what sense they exist at all. There is no solution to this problem in sight.

The emergence debate has already grown quite large. More and more articles and books are published on the different aspects of emergence. The concept of emergence is also used in other fields, for example in cognitive science, in theories of self-organization and in the philosophy of biology, but in a somewhat different sense than in philosophy of mind.

By far the most important source for this work has been Achim Stephan's *Emergenz: von der Unvorhersagbarkeit zur Selbstorganisation* (1999). I believe it is the most comprehensive and thorough account of emergence ever written. Some sections of this thesis, especially in Part I, are in essence merely summaries of sections of Stephan's book. An updated version of the book in English would be more than welcome[1]. Another important source has been the writings of Jaegwon Kim. For more than twenty years, Kim has been clarifying difficult concepts of philosophy of mind like supervenience, reduction, mental causation and lately also emergence.

I had two main objectives writing this work. The first was to make sense of the concept of emergence and the history of emergentism. The other was to find out the significance of emergence for current philosophy of mind. At least in some ways I have reached these objectives, and I hope I can convey some of this to the readers. The two parts of this work correspond to these objectives.

It should be noted that the main focus in this work will be on emergent *properties*, not emergent laws, structures or entities. The latter are only briefly mentioned in the historical overview and some others sections of Part I. The reason for this is that the current debate on emergence is mainly concerned with emergent properties. However, one problem with this approach is that the concept of property in general is extremely problematic. There is no consensus among philosophers on what properties are (universals, particulars, tropes etc.) or whether they even exist. It is not possible to discuss this huge issue in any detail here. Hopefully I will later have the possibility to examine what implications this and other broader issues in metaphysics have on the problem of emergence.

The aim of Part I is to explain what emergence is, what is the history of emergentism, what kinds of theories of emergence there are and what are the most significant aspects of emergence for philosophy of mind. In Chapter 2, I will briefly go through the history of emergentism, focusing on British Emergentism. In Chapter 3, I will discuss the different characteristics of theories of emergence, following closely Stephan's account. Finally, in Chapter 4, I will go through the different types of theories of emergence and their characteristics.

Part II is concerned with emergence in the philosophy of mind. The aim is to show that the concept can be given a well-defined sense and that there truly are

[1] Stephan has already published some articles on emergence in English (see for example 1992, 1998, 2002).

emergent properties in this sense. In Chapter 5, I will try to explain the reasons for the failure of reductive physicalism and clarify some central concepts like reduction and supervenience. Perhaps the most important part of this work is Chapter 6. There I will go through some important arguments for qualia emergentism, describe the functional model of reduction and then present the new qualia emergentism that is based on these arguments and this model. In Chapter 7, I will discuss the problem of mental causation, particularly Kim's supervenience argument and the responses to it.

I would like to thank my supervisor Sami Pihlström for his enlightening comments and helpful assistance and all the Erasmus students[2] in Bremen 2003-2004 for making the time of writing this thesis the time of my life.

[2] Including the Turks.

PART I

What is Emergence?

2. The History of Emergentism

In this chapter I will briefly go through the history of emergentism, focusing on the tradition of British Emergentism[3]. This tradition began with John Stuart Mill (1806-1873) and faded before the middle of the 20th century. The heyday of British Emergentism was in the 1920s, when Samuel Alexander (*Space, Time and Deity*, 1922), C. Lloyd Morgan (*Emergent Evolution*, 1923) and C. D. Broad (*The Mind and its Place in Nature*, 1925) published their main works. (McLaughlin 1992, 49).

There were also emergentists outside this tradition. The American philosopher Roy Wood Sellars (*Evolutionary Naturalism*, 1922) developed his theory of emergence independently of the British Emergentists (Stephan 1999, 4). Pragmatists like William James, John Dewey and G. H. Mead discussed emergence or concepts close to it (El-Hani & Pihlström, 2002a). In addition, there were philosophers before John Stuart Mill who formulated ideas resembling emergentism. Victor Caston (1997) has argued that Aristotle, Galen and certain other philosophers of the classical period were in fact emergentists[4]. Achim Stephan (1999, 99-130) has pointed out several continental philosophers whose theories have emergentist elements: Johann Christian Reil (1759-1813), Hermann Lotze (1817-1881), Gustav Theodor Fechner (1801-1887) and Wilhelm Wundt (1832-1920).

It would be interesting to further discuss this early history of emergentism, but it would require a study of this own. In this overview, I will concentrate on the works of the three most important emergentists: Alexander, Lloyd Morgan and Broad. At the end, I will also go through the emergence theories of Nagel, Popper and Bunge. These three philosophers formulated their theories in the time when British Emergentism had already faded, but the new emergentism had not yet emerged.

It is important to remember that the context in which the theories of Alexander, Lloyd Morgan and Broad were formed was the controversy between mechanism and vitalism that was at its fiercest at the end of the 19th and the beginning of the 20th century. According to vitalism, organic phenomena could not be explained without appealing to non-physical factors that make living beings what they are. The

[3] The expression "British Emergentism" comes from Brian Mclaughlin (1992).
[4] However, for Caston the defining characteristics of emergence are supervenience and downward causation. This is a rather unusual way of understanding emergence. Different definitions of emergence will be discussed in Chapter 4.

most prominent defenders of vitalism were Hans Driesch (1867-1941) and Henri Bergson (1859-1941). The former postulated "entelechies", non-physical elements necessary for the explanation of organic processes, the latter appealed to "élan vital", a driving force responsible for the creation of new forms of life. According to mechanism, everything could be explained mechanically, even organic phenomena. The emergentists rejected both mechanism and vitalism and offered a third alternative. According to emergentism, all beings and structures, whether living or non-living, are composed of the same basic elements, but there are nonetheless relevant differences between physical, biological and mental phenomena, and different kinds of explanations must be applied to them. (Stephan 1999, 6-13). Nowadays the mechanism-vitalism controversy is merely a somewhat amusing chapter in the history of philosophy, but emergentism lives on.

2.1. John Stuart Mill: *The System of Logic*

The custom of tracing the history of emergentism back to John Stuart Mill's *The System of Logic* probably stems from Lloyd Morgan. In *Emergent Evolution* he writes:

> The concept of emergence was dealt with (to go no further back) by J.S. Mill in his *Logic* (Bk. III. ch. vi. §2) under the discussion of "heteropathic laws" in causation. The word "emergent", as contrasted with "resultant", was suggested by G. H. Lewes in his *Problems of Life and Mind* (Vol. II. Prob. V. ch. iii. p. 412). (1923, 2-3).

The term "heteropathic" requires some explanation. According to Mill, "homopathic" laws and effects follow the principle of Composition of Causes. This principle states that the joint effect of several causes is identical to the sum of their separate effects (p. 371). The paradigmatic example of this is the addition of forces in mechanics, where also the name for the principle comes from:

> If a body is propelled in two directions by two forces, one tending to drive it to the north and the other to the east, it is caused to move in a given time exactly as far in both directions as the two forces would separately have carried it; and is left precisely where it would have arrived if it had been acted upon first by one of the two forces, and afterwards by the other. This law of

> nature is called, in dynamics, the principle of the Composition of Forces: and in imitation of that well-chosen expression, I shall give the name of the Composition of Causes to the principle which is exemplified in all cases in which the joint effect of several causes is identical with the sum of their separate effects. (1843, 370-371).

According to Mill, most of the effects and laws in nature follow the principle Composition of Causes. Sometimes, however, it is breached. For example, "[t]he chemical combination of two substances produces, as is well known, a third substance with properties different from those of either of the two substances separately, or of both of them taken together" (p. 371). The product of a chemical reaction can in no way be seen as a sum of the reactants. With regard to organic phenomena, the breach is even clearer: "the phenomena of life […] bear no analogy to any of the effects which would be produced by the action of the component substances considered as mere physical agents" (p. 371). Mill calls the laws that cover these kinds of instances *heteropathic* laws (p. 374).

The term "emergence" comes from Mill's student and follower George Henry Lewes (1817-1878). Lewes calls the effects that follow the principle of Composition of Causes "resultants". Mill's heteropathic effects he calls *emergents*[5]. As Brian McLaughlin (1992, 65) has noted, the introduction of this term is Lewes's main contribution to emergentism, but it is an important one, as the word "emergent" or "emergence" has considerably more power in it than "heteropathic law" or "heteropathic effect".

This is how Mill's significance for emergentism is usually seen: he introduced the distinction between homopathic and heteropathic laws and effects, and later Lewes named heteropathic effects emergents. However, Achim Stephan (1999, 87-98) has argued that Mill's influence was much broader.

In the fourteenth chapter of the third book of *A System of Logic*, Mill distinguishes between two kinds of laws: *ultimate* and *derivative*. Derivative laws are those that can be deduced from more general ones, ultimate laws are those that cannot (p. 484). According to Mill, there are fundamental limitations to the possibilities of explanation, as certain kinds of laws are in principle ultimate – namely the laws that connect physical phenomena with conscious experiences:

---

[5] The words "emergence", "emergent" and the like are derived from the Latin verb *emergo* – to arise, to come forth.

> [T]he ultimate Laws of Nature cannot possibly be less numerous than the distinguishable sensations or other feelings of our nature;- those, I mean, which are distinguishable from one another in quality, and not merely in quantity or degree. For example; since there is a phenomenon *sui generis*, called colour, which our consciousness testifies to be not a particular degree of some other phenomenon, as heat or odour or motion, but intrinsically unlike all others, it follows that there are ultimate laws of colour; that though the facts of colour may admit of explanation, they never can be explained from laws of heat or odour alone, or of motion alone, but that however far the explanation may be carried, there will always remain in a law of colour. I do not mean that it might not possibly be shown that some other phenomenon, some chemical or mechanical action for example, invariably precedes, and is the cause of, every phenomena of colour. But though this, if proved, would be an important extension of our knowledge of nature, it would not explain how or why a motion, or a chemical action, can produce a sensation of colour […] (p. 485).

Anyone acquainted with the current qualia debate should find this kind of argumentation eerily familiar. Mill is clearly arguing for the irreducibility of qualia here, more than hundred years before the qualia debate took off. (Stephan 1999, 89-90).

Mill also distinguishes between ultimate and derivative *properties*. Ultimate properties belong to elementary substances unconditionally, they are not dependent on any other factors or causes. Because all properties of complex systems depend on certain factors (i.e. a certain microstructure), they are in Mill's terminology derivative properties. (P. 581; Stephan 1999, 84-85).

Mill's approach was quite different from the later British Emergentists: his aim was to bring the scientific knowledge gathered in the run of the centuries back to its roots, and he claimed these roots were experience and induction. In the course of this enterprise he formulated ideas that more than sixty years later were used by the emergentists for something that Mill never had in mind: for trying to show a third way between mechanism and vitalism, for explaining the appearance of new entities and properties in the course of evolution, or for taking a metaphysical position in the mind-body problem. (Stephan 1999, 97)

Most clearly the influence of John Stuart Mill can be seen in C. D. Broad's theory of emergence. Among other things, Broad distinguishes between ultimate and derivative laws, refines Mill's distinction of ultimate and derivative properties, and argues like Mill that for every distinguishable phenomenal experience there must exist an ultimate law. (Stephan 1999, 98).

2.2. Alexander and Lloyd Morgan

In *Space, Time and Deity* (1920), Samuel Alexander presents a comprehensive metaphysical system, in which the concept of emergence plays an important role. According to Alexander, at its lowest level the world consists of basic space-time, from which higher levels of existence emerge. Roughly speaking, from space-time emerges matter, from matter the secondary qualities and life, from life consciousness or mind, and finally, from mind the quality of Deity. (1920, 45-70). Let us take a closer look on what Alexander means with emergence:

> [A]s in the course of Time new complexity of motions comes into existence, a new quality emerges, that is, a new complex possesses as a matter of observed empirical fact a new or emergent quality. […] The emergence of a new quality from any level of existence means that at that level there comes into being a certain constellation or collocation of motions belonging to that level, and possessing the quality appropriate to it, and this collocation possesses a new quality distinctive of the higher complex. The quality and the constellation to which it belongs are at once new and expressible without residue in terms of the processes proper to the level from which they emerge […]. (1920, ii, 45)

> The higher quality emerges from the lower level of existence and has its roots therein, but it emerges therefrom, and it does not belong to that lower level, but constitutes its possessor a new order of existent with its special laws of behaviour. The existence of emergent qualities thus described is something to be noted, as some would say, under the compulsion of brute empirical fact, or, as I should prefer to say in less harsh terms, to be accepted with the "natural piety" of the investigator. It admits no explanation. (1920, ii, 46-47).

This may seem somewhat confusing at first. Presumably something like the following is intended: when at a certain level something emerges, there comes into being a new structure with a new quality. The quality is novel in the sense that it has not occurred before and it could not have been predicted beforehand. The existence of this quality cannot be further explained, it must be accepted with "natural piety". Therefore, the key characteristics of emergent qualities are novelty and unpredictability.

However, there appears to be a conflict in Alexander's theory (Stephan 1999, 47-52). According to Alexander, the new quality is "expressible without residue in terms of the processes proper to the level from which they emerge" (1920, ii, 45). On the other hand, Alexander is committed to a strong determinism, where a Laplacian demon given the physical state of the universe at a certain instant or instants could calculate the physical state of the universe at any later instant (see for example 1920, ii, 73). If the emergent qualities can be expressed without residue in terms of the lower levels, they can in the end be expressed in terms of physics (or in terms of pure motions in space-time, as Alexander would put it), and there seems to be no reason why a Laplacian demon could not predict them.

Achim Stephan (1999, 48-50) has showed that Alexander's system can be saved with a minor modification – i.e., by abandoning the requirement that emergent qualities must be expressible without residue in terms of the lower level. With this modification, we would have the following picture: when a novel structure with an emergent quality comes into being, the behaviour of the components of the structure and the structure itself can be explained "without residue" in terms of the lower levels. Therefore, they can also be predicted before they first appear. However, the *emergent quality* that this structure has is in principle unpredictable and cannot be further explained.

I do not believe that it is worthwhile to delve deeper into the rather obscure world of Alexander in this context. Let us next look briefly at Lloyd Morgan and then turn to C. D. Broad, whose work has by far the most significance for the contemporary discussion of emergence.

In *Emergent Evolution* (1923) Lloyd Morgan gives his own account of emergence, building on the theories of Mill and Lewes. Following Mill, he gives an example from chemistry as a typical case of emergence: "When carbon having certain properties combines with sulphur having other properties there is formed, not a mere mixture but a new compound, some of the properties of which are quite different from those of either component." (P. 3). As for Alexander, for Lloyd Morgan the defining characteristics of emergent qualities are novelty and unpredictability:

> The point of emphasis […] is this. Let there be three successive levels of natural events, A, B, and C. Let there be in B a *kind of relation* which is not present in A; and in C a kind of relation, not yet present in B or in A. If then one lived and gained experience on the B-level,

> one could not predict the emergent characters of the C-level, because the relations, of which they are the expression, are not yet in being. Nor if one lived on the A-level could one predict the emergent character of b-events, because *ex hypothesi*, there are *no such events* as yet in existence. What, it is claimed, one cannot predict, then, is the emergent expression of some new kind of relatedness among pre-existent events. One could not foretell the emergent character of vital events from the fullest possible knowledge of physico-chemical events only […] (Pp. 5-6).

Lloyd Morgan stresses the point that the emergent theory is naturalistic in the sense that it does not invoke any extra-natural powers like entelechies or God (p. 2). It is not, however, mechanistic. Mechanistic theories try to explain everything in terms of resultant effects that are in principle calculable. The point of emergentism is that this kind of explanation is inadequate. "Resultants there are; but there is emergence also." (P. 8).

According to Lloyd Morgan, an emergent quality (or "a new kind of relatedness", as he sometimes puts it) makes a difference to the way things "run their course" at lower levels. In other words, the manner in which lower-level events happen "depends on" the new kind of relatedness. This must be accepted with Alexander's "natural piety". (Pp. 15-18). Therefore, in contrast to Alexander, Lloyd Morgan is clearly and strongly committed to downward causation (see sections 3.7 and 7.2).

2.3. C. D. Broad: *The Mind and Its Place in Nature*

The peak of British Emergentism and in some ways the basis for the current emergence debate is C. D. Broads *The Mind and It Place in Nature* (1925). It is based on the course of Tarner lectures that Broad gave in Cambridge in 1923. The purpose of these lectures was to demonstrate "the relation or lack of relation between the various sciences". It is curious that apparently Broad himself did not value *The Mind and Its Place in Nature* very high – for example, in his "Autobiography" (1959) it is hardly mentioned at all.

In the second chapter of *The Mind and Its Place in Nature*, "Mechanism and Its Alternatives", Broad discusses different ways of accounting for the characteristic

differences among objects (p. 48). A particular case of this general problem is the controversy between mechanism and vitalism.

Broad distinguishes three possible types of theory that account for the characteristic differences of behaviour. First, there are theories that hold that "the characteristic behaviour of a certain object or class of objects is in part dependent on the presence of a peculiar *component* which does not occur in anything that does not behave in this way" (p. 55). An example of this kind of theory is *substantial vitalism*, which assumes that a necessary factor in explaining the behaviour of living objects is the presence of an "entelechy", a peculiar component which does not occur in inorganic matter. According to Broad, this is also the view commonly taken about chemical behaviour: different chemical compounds behave differently because of different components. Theories of this kind need not deny that differences of structure are also relevant to explaining behaviour: for example, two chemical compounds with identical components can behave differently because of different microscopic structure. (Pp. 55-58).

The other two kinds of theories deny that peculiar components are necessary for explaining behaviour, and try to explain the differences wholly in terms of difference of structure. The first of these theories holds that "the characteristic behaviour of the whole *could* not, even in theory, be deduced from the most complete knowledge of the behaviour of its components, taken separately or in other combinations, and of their proportions and arrangements in the whole" (p. 59). Broad calls this the "theory of emergence", and I will soon discuss it in greater detail.

According to the third kind of theory the behaviour of a whole *could*, at least in theory, be *deduced* from a sufficient knowledge of the behaviour of the components. This kind of theory Broad calls "mechanistic". The most obvious example of a class of objects to which mechanistic theories apply is mechanical devices. For example, there is no doubt that the behaviour of a clock can be deduced from sufficient knowledge of its components. (Pp. 59-60). Contemporary reductive physicalism would probably be a mechanistic theory in Broad's terminology.

Now let us return to Broad's theory of emergence. Its core is the following definition that will play a central role in the discussion of emergence to follow:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of

> constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). (P. 61).

In other words, a property of a whole is emergent if and only if it cannot be deduced from the most complete knowledge of the properties of its components in isolation or in other wholes. But why do we need the lengthy expression "in isolation or in other wholes"? Is it not enough just to say "the properties of A, B, C"? The most likely reason why Broad is using the longer expression is that he has foreseen a certain argument that would show that there can be no emergent properties[6]. (Stephan 1999, 32-35).

According to this argument, one of the kinds of properties that an object can have is that under certain conditions it becomes a part of a compound with certain properties. For example, one of the properties of silver is that under certain conditions it combines with chlorine to form a compound with the properties of silver-chloride. Thus the properties of silver-chloride can be deduced from the properties of silver, and they cannot be emergent. The same applies for all supposedly emergent qualities. Therefore we need to somehow limit the properties that can be used in the deductions, and presumably this is the aim of Broad's expression "in isolation or in other wholes". If only the properties that the components have been observed to manifest in isolation or in other compounds can be used in deducing the properties of the whole, the problem can be avoided. (Stephan 1999, 32-35).

According to Broad, a law that connects an emergent property of a structure with the properties of the components of the structure is a *unique*, *ultimate* and *irreducible* law. This means that it is not a special case of a more general law and that it does not arise from a combination of more general laws. It is a law that could have been discovered only by studying this particular case. (P. 65). This corresponds to Mill's idea of ultimate laws (see section 2.1.). When I later use the expression "emergent law", I am referring to laws of this kind.

[6] This interpretation is supported by Broad's remarks on pages 65-66.

Now let us turn to Broad's archangel example. In the section "Mechanistic Theories" Broad is discussing the theoretical limitations of prediction. To get rid of practical mathematical difficulties, he proposes that we replace Sir Ernest Rutherford by a mathematical archangel. If the emergent theory of chemical compounds is true, "a mathematical archangel, gifted with the further power of perceiving the microscopic structure of atoms as easily as we can perceive hay-stacks, could no more predict the behaviour of silver or of chlorine or the properties of silver-chloride without having observed samples of those substances than we can at present" (p. 71).

Furthermore, even if the mechanistic theory of chemistry were true (instead of the emergent theory), there would be a theoretical limit to the deduction of the properties of chemical elements and compounds:

> Take any ordinary statement, such as we find in chemistry books; e.g., "Nitrogen and Hydrogen combine when an electric discharge is passed through a mixture of the two. The resulting compound contains three atoms of Hydrogen to one of Nitrogen; it is a gas readily soluble in water, and possessed of a pungent and characteristic smell." If the mechanistic theory be true the archangel could deduce from his knowledge of the microscopic structure of atoms all these facts but the last. He would know exactly what the microscopic structure of ammonia must be; but he would be totally unable to predict that a substance with this structure must smell as ammonia does when it gets into the human nose. The utmost that he could predict on this subject would be that certain changes would take place in the mucous membrane, the olfactory nerves and so on. But he could not possibly know that these changes would be accompanied by the appearance of a smell in general or of the peculiar smell of ammonia in particular, unless someone told him so or he had smelled it for himself. If the existence of the so-called "secondary qualities," or the fact of their appearance, depends on the microscopic movements and arrangements of material particles which do not have these qualities themselves, then the laws of this dependence are certainly of the emergent type. (Pp. 71-72)

The quote is long, but this passage is of great importance. Among other things, it is perhaps the earliest formulation of the "knowledge argument" that has been later advocated by Thomas Nagel (1974) and Frank Jackson (1982, 1986). The point of the knowledge argument is to show that even if someone knew everything about completed physical sciences, he/she wouldn't necessarily know everything there is to know. For example, if he/she had lived in a black-and-white environment, he/she

wouldn't know what it is like to see red or other colors. Therefore, physical sciences cannot give the whole truth of the world.

This also seems to be the point of Broad's archangel example[7]. According to Broad, it is not *a priori* impossible that chemistry and biology were mechanical, but they cannot be the whole truth of the material world, because smells, tastes, colours and other secondary qualities cannot be mechanically explained. The laws connecting microscopic particles or events with secondary qualities must be emergent laws, "[a]nd no complete account of the external world can ignore these laws". (P. 72)

According to Broad, mechanism would introduce a sort of a unity into the world and the sciences, but on the emergent theory, the world and the sciences that deal with it form at best a kind of hierarchy. An emergentist need not deny that the world is ultimately composed of a single kind of stuff, but he must recognize that within this stuff there are aggregates of different orders. Therefore, there are also two fundamentally different types of law: "trans-ordinal" and "intra-ordinal". Trans-ordinal laws connect the properties of aggregates of adjacent orders, intra-ordinal laws the properties of aggregates of the same order. (Pp. 76-78)

It seems that Broad thought that all trans-ordinal laws are also emergent laws. He never says this explicitly, but writes for example: "a trans-ordinal law would be a statement of the irreducible fact that an aggregate composed of aggregates of the next lower order in such and such proportions and arrangements has such and such characteristic and non-deducible properties" (p. 78). This would make trans-ordinal laws by definition emergent laws. In any case, at least all "trans-physical" laws are necessarily emergent. These are laws that connect properties of aggregates with secondary qualities, like smells and tastes and colours. (Pp. 79-80). Here Broad comes again close to the ideas of John Stuart Mill (see the quote from Mill on page 10).

Broad also distinguishes between three kinds of properties: (i) "ultimate properties" are properties of a certain order that all aggregates of this order but no aggregate of lower orders possess, and which could not be deduced from the structure of the aggregate and the properties of its constituents, (ii) "reducible properties" are properties which are characteristic of a certain order, but which could in theory be deduced from the structure of the aggregate and the properties of its constituents, (iii) "ordinally neutral properties" are properties that an aggregate of a certain order shares

[7] There are, however, significant differences between Broad's archangel example and later versions of the knowledge argument. This subject will be further discussed in section 6.1.

with aggregates of lower orders. (P. 78). Ultimate properties are clearly emergent properties.

2.4. The Fall of British Emergentism

The theories of the British Emergentists were widely discussed and heavily critized in the 1920s. Emergence was the main theme of several symposiums and the main works of the emergentists received a lot of attention in philosophical journals. (Stephan 1999, 131). However, it seems that philosophical criticism was *not* the reason that lead to the fall of British Emergentism. Achim Stephan (1999, 129-155) has analyzed the different kinds of critique aimed at emergentism and showed that they did not pose a significant threat to it. No fundamental inconsistency or philosophical error was pointed out (see also McLaughlin 1992 and Kim 1999, 3-4).

Brian McLaughlin (1992) has argued that the main reason for the downfall was advances in science. Soon after Broad's *The Mind and Its Place in Nature* (1925) was published, quantum mechanics was discovered, which made possible (among many other things) the explanation of chemical bonding in terms of electro-magnetism. This in turn led to the development of molecular biology and eventually to the discovery of DNA. These advances made the existence of emergent properties or laws in chemistry and biology highly implausible. I have to point out that they did in no way affect the plausibility of emergence with regard to philosophy of mind. Chemistry and biology, however, were the main examples of emergence for most of the British Emergentists.

Another reason for the fall was the rise of logical positivism in the 1930s (Kim 1999, 3-4). This trend in philosophy was anti-metaphysical and hyper-empiricist and the supposedly vague concept of emergence had no place in its view of the sciences. Some positivists tried to adapt the concept of emergence for their own purposes, and for example Ernest Nagel (see the next section) gave a theory-relative interpretation for it.

I believe these two were the main reasons for the fall of emergentism. It was not philosophically refuted and the tradition did not die out, it merely lost its place in the centre of the field of philosophy. Articles on the works of British Emergentists were published in journals until the 1960s, and by that time the concept of emergence was already finding its way to the philosophy of mind (Stephan 1999, 131, 159). The

failure of reductionism lead to the rise of “nonreductive physicalism”, and it can be argued that this actually amounted to the return of emergentism. If this is so, it is no exaggeration to say that we have been under the reign of emergentism since the early 1970s. (Kim 1999, 4-5).

2.5. Nagel, Popper and Bunge

Before the concept of emergence returned to the mainstream of philosophy, some prominent philosophers developed their own theories of emergence. Perhaps the most significant ones of these were the theories of Ernest Nagel, Karl Popper and Mario Bunge. I will begin with Ernest Nagel’s critical and influential views on emergence that actually do not form a full-blown theory, but are rather critical revisions of the ideas of the British Emergentists.

In his classic work *The Structure of Science* (1960), Ernest Nagel critically examines the concept of emergence in the spirit of logical positivism and presents a theory-relative interpretation of it. According to Nagel, emergence can be formulated in two ways: as a thesis about the nonpredictability of certain characteristics of things, or as an evolutionary thesis according to which genuinely novel traits and structures repeatedly come into being (pp. 366-367). Because the latter form has little to do with philosophy of mind, I will only discuss the first form here.

According to this first form, there are at least some properties of objects that are impossible to predict from the most complete knowledge of the properties and the order of the parts of the object. This closely corresponds to Broad’s idea of emergence (see section 2.3). However, according to Nagel, this concept is theory-relative, not absolute. He gives the following reasons for this.

First of all, the deduction of properties is always impossible, because it is not properties but statements or propositions which can be deduced. In addition, statements about properties of complex wholes can be deduced from the statements about their constituents only with the help of an appropriate theory concerning these constituents. All the terms of the deducible statement must occur in this theory or in the assumptions adjoined to it in the particular case at hand. For example, the statement “Water is translucent” cannot be deduced from the theories concerning hydrogen and oxygen if the terms “water” and “translucent” are not included in these

theories. (Pp. 367-369). Therefore, whether a certain property is emergent or not depends on the theories we have at hand. This analysis renders the concept of emergence purely theory-relative:

> It is clear, therefore, that to say of a given property that it is an "emergent" is to attribute to it a character which the property may possess relative to one theory or body of assumptions but may not possess relative to some other theory. Accordingly, the doctrine of emergence (in the sense now under discussion) must be understood as stating certain *logical* facts about formal relations between statements rather than any experimental or even "metaphysical" facts about some allegedly "inherent" traits of *properties* of objects. (p. 369).

Similar critical views on emergence had already been put forward by various philosophers, most importantly by Hempel and Oppenheim (1948). The point was that properties are not emergent in any absolute sense, only relative to a certain theory, and increase in knowledge and new theories can make a property that once was emergent no longer emergent (and vice versa). Therefore, emergence is not an ontological concept, but merely epistemic. Nagel also claims that there are countless examples of properties that are emergent in this sense, and there is nothing special about the properties that the emegentists usually claim to be emergent (pp. 372-374).

However, an important part of Broad's emergentism and the central claim of the current proponents of emergence is that some properties, namely phenomenal properties (or secondary qualities, or qualia), are *necessarily* emergent so that *no* theory can explain them and increase in knowledge will not change this. This is because phenomenal properties cannot be functionally or behaviourally analyzed, and thus they are necessarily irreducible. Another way of putting this is that there always remains an *explanatory gap* between the mental and the physical: phenomenal properties cannot be explained in physical terms (Levine 1983, more on this in section 6.1). If this is true, there are some properties that are emergent in an absolute, not theory-relative sense. (Stephan 1999, 145-146).

Now I will turn to the emergence theory of Karl Popper[8] that he has put forward in *The Self and Its Brain* (1977). It is an important part of his peculiar

[8] This summary of the theories of Popper and Bunge is somewhat based on Stephan (1999, 178-185).

metaphysical system, in which the universe is divided into three different but interacting "worlds" – the physical world (stones, stars, plants, animals, etc.), the mental world (feelings, thoughts, perceptions, etc.) and the world of products of the human mind (languages, stories, aeroplanes, theories, mathematical constructions, etc.). According to Popper, emergence means "the fact that in the course of evolution new things and events occur, with unexpected and indeed unpredictable properties; things and events that are new, more or less in the sense in which a great work of art may be described as new" (p. 22). In the evolution of the universe, at least the following stages can be distinguished: 1. The production of heavier elements. 2. The emergence of life. 3. The emergence of sentience. 4. The emergence of the consciousness of self and death. 5. The emergence of the human language, together with theories of the self and death. 6. The emergence of works of art and science. (P. 16).

For Popper, the main arguments for emergence are the indeterministic nature of the universe that has been revealed by quantum mechanics and the existence of downward causation. He claims that the critique against emergence has come from three sides: 1. From the determinists, who claim that there are no objectively chancelike events or probabilities. 2. From the classical atomists, who claim that all physical bodies and all organisms are nothing but structures of atoms. 3. From the upholders of a theory of potentialities, who claim that if something seems to emerge, it must have already existed as a "disposition" or "potentiality" of the physical particles or structures involved. He then argues that the arguments 1. and 3. are based on classical physics and quantum mechanics has made them obsolete. Against argument 2. he points out that new atomic arrangements may lead to properties that are not derivable from the statements describing the arrangement of the atoms. (Pp. 22-31).

However, what is most interesting is that the arguments 1.-3. are *not* typical arguments against emergentism and are completely beside the point. Popper would have noticed this if he had examined the works of the early emergentists more closely: Determinism is one of the characteristics of classic theories of emergence; the British emergentists believed that all structures are composed of atoms and nothing else; and for example Broad took into account the possibility that emergent properties are dispositions or potentialities (see 1925, 65-67). Therefore, this part of Popper's discussion is largely irrelevant for the problem of emergence.

Downward causation forms an important part of Popper's theory of emergence. For Popper, downward causation means the fact that macro structures as a whole act upon their components (p. 19). Popper sees downward causation everywhere: "every single arrangement of negative feedback, such as a steam engine governor, is a macroscopic structure that regulates lower level events" (p. 19); "[d]ownward causation is of course important in all tools and machines which are designed for some purpose" (p. 19); "[t]he most interesting examples of downward causation are to be found in organisms and their ecological systems, and in societies of organisms" (p. 20). Unfortunately, Popper does not even try to explain *how* the macro structures act upon their components. The hardest part of the problem of downward causation is finding a plausible mechanism of macro-to-micro causation, and in Popper's theory, no such mechanism can be found.

In general, Popper's theory of emergence is vague and confused. He leaves some of the central concepts like "unpredictability" undefined. He sees emergence and downward causation everywhere, but ignores the most difficult problems related to them or is not even conscious of them. Popper's emergentism does not even come close to the clarity and refinement of the theories of some of the British Emergentists. (Stephan 1999, 182).

In contrast to Popper, Mario Bunge (1977) has presented a theory of emergence that involves an extremely *weak* interpretation of the concept. Bunge characterizes emergent properties as follows:

> Temperature and entropy are properties of an atomic aggregate, not possessed by any of its atomic components. Likewise the capacity to self-duplicate is a property of deoxyribonucleic acid molecules that none of their components (i. e. the nucleotides) possess. These are examples of emergent properties, or properties characterizing a system as a whole and which the system components do not have. (1977, 502).

It is clear that Bunge is characterizing *systemic properties* here (see section 3.2). In fact, he does not make any distinction between systemic and emergent properties. This is made explicit in the following definition:

> Let *P* be a property of a complex thing *x* other than the composition of *x*. Then (i) *P* is resultant or hereditary if *P* is a property of some components of *x*; (ii) otherwise, i.e. if no component of *x* possesses *P*, *P* is *emergent*, *collective*, *systemic*, or *gestalt*. (1977, 502).

However, already the early emergentists distinguished between *collective* (systemic) and *emergent* properties. Collective properties are properties of a system which none of the components have. Some collective properties are irreducible, and these are called emergent properties. According to Bunge, there are no irreducible collective properties, and thus no emergent properties in this sense: "We recognize the fact of emergence but assume that every emergent can be accounted for in terms of a system's components and the couplings among them" (1977, 503). This shows that Bunge's view on emergence differs radically from that of the British Emergentists and the current emergentists.

Bunge's concept of emergence is also *relative*: "Thus the ability to think is an emergent property of the primate brain relative to its component neurons, but it is a resultant property of the primate because it is possessed by one of the latter's components, namely its brain" (1977, 502).

Bunge calls his theory *rational emergentism*. It is "a doctrine differing from both the irrationalist emergentism of the holists and the rationalist flattening (or leveling) by the mechanists, energetists and idealists" (1977, 503). Two postulates constitute the kernel of this doctrine. The *emergence postulate* states that some of the properties of every system are emergent. The *rationality postulate* states that every emergent property of a system can be explained in terms of properties of its components and of the couplings amongst these. (1977, 503).

Rational emergentism is supposed to provide a middle way between irrational emergentism and reductionism. However, it differs from reductionism only if we employ an incredibly strong concept of reduction. For example, Bunge's "emergentism" is perfectly compatible with reductionism based on the Nagelian model of reduction or the new functional model of reduction (see sections 5.2 and 6.3). On the other hand, it is quite rough to imply that the elaborate theories of C. D. Broad and the others were "irrational".

It would be more appropriate to see Bunge's theory as a form of reductive physicalism and reserve the term emergentism for theories with a stronger concept of emergence. For the current discussion of emergence, Bunge's theory is rather

insignificant – mental properties and many other properties are trivially emergent in Bunge's sense. The problem is whether mental properties are emergent in the sense of being fundamentally irreducible.

3. The Central Characteristics of Theories of Emergence

In this chapter I will go through the central characteristics (Merkmale) of emergentist theories, following closely Achim Stephan's (1999) thorough account. These characteristics are *naturalism*, *systemic properties*, *novelty*, *hierarchy of levels of existence*, *diachronic* and *synchronic determinism*, *irreducibility*, *unpredicability* and *downward causation*. They primarily characterize the theories of the British Emergentists, but explaining them carefully should also lay a good foundation for the discussion to follow. I will focus on the doctrines that have the most significance for contemporary philosophy of mind.

3.1. Naturalism

The first characteristic of emergentism is the acceptance of a fully *naturalistic* standpoint. What is meant by this is that all *supernatural* explanations are unacceptable. It is not allowed to appeal to Cartesian souls, God, entelechies, faeries, élan vital or anything like that. There are no supernatural entities or forces. On the contrary, all structures in nature are composed of the same basic elements. There are no special components that only living or conscious beings would have. All properties are instantiated by structures composed of natural elements. (Stephan 1999, 14-15). For example, Lloyd Morgan writes: "The naturalistic contention is that, on the evidence, not only atoms and molecules, but organisms and minds are susceptible of treatment by scientific methods fundamentally of like kind; that all belong to one tissue of events; and that all exemplify one foundational plan." (1923, 2).

With the acceptance of naturalism an emergentist can hold on to an empiristic and scientific world-view without falling into reductionism. However, it must be noted that there are several different forms of naturalism and the concept is quite problematic. For example, according to *strong* naturalism, all mental properties can be "naturalized", that is, reduced to physical properties and phenomena. It would be somewhat less problematic to say that emergentists are committed to *physical monism*, according to which all entities in the world are composed of physical elements.

Therefore, all properties, including emergent properties, are instantiated by physical structures. (Stephan 1999, 15-16, 66-67; Schmitt 1995).

3.2. Novelty and Systemic Properties

Another characteristic of theories of emergence is the idea that in the course of time, there repeatedly comes into being something *genuinely novel*. For example, according to Lloyd Morgan (1923, 1) "the orderly sequence, historically viewed, appears to present, from time to time, something genuinely new. Under what I here call emergent evolution stress is laid on this incoming of the new. Salient examples are afforded in the advent of life, in the advent of mind, and in the advent of reflective thought."

An important question here is: what exactly is meant by "something new"? New structures, new entities, new properties or even new laws? What is certainly *not* intended is the coming into being of *numerically new* entities or properties. For example, the property of having the weight 459,0924707896914 g would not be called genuinely novel, even if it were instantiated for the first time. What is meant is instantiations of a type that has not been instantiated before. This can mean a new type of structure or a new type of property. A law can also be new only in the sense that it has not been instantiated before – it would be odd to claim that new laws of nature come into being.[9] However, what is usually meant by "emergent laws" is irreducible or ultimate laws. (Stephan 1999, 17-18).

The concept of novelty is problematic and should be more thoroughly discussed, but it is not that relevant for philosophy of mind. Therefore, in short and roughly speaking, the characteristic *novelty* refers to new types of structures or new types of properties.

If we assume that systems with identical microstructures cannot have different properties (*mereological supervenience*, see section 5.3), it follows that a necessary condition for the appearance of a novel property is the appearance of a novel structure. If the structure would not be new, the property would have been already istantiated.

[9] This of course depends on what we take laws to be, which is another huge issue that I don't want to discuss here.

This idea is widely accepted in current philosophy and the British Emergentists were clearly committed to it. (Stephan 1999, 19-20).

The idea of genuine novelty presupposes a certain basic distinction. We must distinguish between two kinds of properties that complex entities can have: properties that also some of the components have, and properties that none of the components have. The latter kinds of properties are called *systemic properties*[10]. It is obvious that only systemic properties can be genuinely novel. (Stephan 1999, 20).

Both of these theses – systemic properties and genuine novelty – are hard to deny. If there were no systemic properties, all properties of structures would be properties that also some of the components have, but countless examples prove this wrong. To claim that there is no genuine novelty would be to claim that all types of systems and properties that now exist have always existed. To say the least of it, this claim has no empirical support. (Stephan 1999, 21).

3.3. Hierarchy of the Levels of Existence

Another typical feature of theories of emergence is the layered view of nature. On this view, all things in nature belong to a certain level of existence, each according to its characteristic properties. These levels of existence constitute a hierarchy of increasing complexity that also corresponds to their order of appearance in the course of evolution. The fundamental level is the level of physics, and for each higher level there is a special science that deals with it. Different emergentists postulate a different number of levels, but the most central ones are the level of inorganic or material things, the level of life and the level of mind or consciousness. (O'Connor & Wong 2002, Stephan 1999, 23).

According to Jaegwon Kim (1998, 15-19), a layered view of the world like this has played an important role not only in the theories of the British Emergentists but also in the metaphysics and philosophy of science of the 20th century in general. The fundamental relation that creates this hierarchical model is the part-whole relation: all entities of a given level are entirely composed of entities of lower levels, except for the entities of the basic level, which have no structure. Sometimes this model is

[10] They have also been called *collective*, *gestalt*, *structural* or even *emergent* properties.

expressed in terms of concepts and languages rather than of entities and their properties. This layered view of the world has been at the backround of debates on many issues, for example reductionism, the mind-body problem, the role of special sciences, and so on.

3.4. Synchronic and diachronic determinism

Two important characteristics of classic emergentism are *determinism* and *unpredictability*. In the time of British Emergentism it was often assumed that these two are incompatible: the world must be either indeterministic and unpredictable or deterministic and predictable. The emergentists, however, argued that the world can be both deterministic and unpredictable at the same time. This view is nowadays generally accepted. Unpredictability is an epistemological concept, determinism ontological. It is possible that there are fundamental limitiations to the possibilities of prediction, so that even deterministic phenomena can not be predicted, even in principle. (Stephan 1999, 26).

It is important to distinguish between two kinds of determinism: *synchronic* and *diachronic*. The idea of diachronic determinism is what is usually associated with the word determinism: roughly speaking, that under the same initial conditions the same events will happen and the same structures will appear. Synchronic determinism[11] means that the properties and behavioural dispositions of a system are nomologically dependent on its microstructure (i.e. the properties and order of the components). There can be no difference in the properties of the system without there being a difference in microstructure. (Stephan 1999, 26). This is a form of what is nowadays called *supervenience*. The concept of supervenience has been widely discussed, and there are different formal definitions for it. Supervenience will be further discussed in section 5.3.

Synchronic determinism should be accepted by all philosophers with a naturalistic standpoint. For example, it would be possible to claim that there can be two microstructurally identical systems with different properties, but then one would

[11] It would perhaps be more appropriate to talk of synchronic *determination*. However, Stephan draws an analogy between synchronic and diachronic "determinism" (Determiniertheit), and to reflect this, I have used the word "determinism" for both.

have to explain this difference in terms of some supernatural factors, or accept that properties can change and differ without a specific reason. (Stephan 1999, 27).

Synchronic determinism is also evident in the works of the British Emergentists. The clearest formulation is in the passage of C. D. Broad's *The Mind and Its Place in Nature* that we are already familiar with:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties[…](p. 61).

Broad clearly states here that all systems with identical microstructures have the same characteristic properties, that is, the properties of a system are determined by the microstructure of the system.

Diachronic determinism is central for evolutionally oriented theories of emergence, but has little significance for the discussion of emergence in contemporary philosophy of mind. In modern terminology it could be formulated as follows: it is not possible that different structures come into being in two worlds with the same initial conditions and the same laws of nature. This characteristic can most clearly be seen in the works of Lloyd Morgan and Alexander. For example, Lloyd Morgan writes: "[I]f there be a natural plan of emergence, then every effect is strictly determinate in accordance with the nature of that plan; […] novelty itself is thus caught up in the web of causal nexus" (1923, 282). And Alexander: "A calculator given the state of the universe at a certain number of instants or at one instant with the laws of its change could, given sufficient powers, calculate what the spatio-temporal condition of the world would be at any given later instant" (1920, ii, 73).

3.5. Irreducibility

The next two characteristics are irreducibility and unpredictability. There is a connection between these two: systemic properties that are irreducible are also in principle unpredictable before their first appearance. However, it is possible that a property is reducible, but for some other reason unpredictable. The concept of

unpredictability is in this sense more complex than irreducibility, and so irreducibility will be discussed first. (Stephan 1999, 32).

Irreducibility is a central concept in the contemporary discussion of emergence. It also plays a central role in Broad's theory of emergence. The already familiar passage from *The Mind And Its Place in Nature* may serve as a starting point here:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). (1925, 61).

One problem with this definition is that strictly speaking no properties can be deduced, only laws or law statements[12]. In modern terminology, and taking this problem into account, the thesis of irreducibility could be formulated as follows (Stephan 1999, 36):

> *Irreducibility*. Let there be a system *S* whose systemic property *E* is nomologically dependent on the microstructure $<c_1, \ldots, c_n \mid o>$ of *S* (i.e. the components $c_1, \ldots, c_n$ and their specific order *o*). *E* is *irreducible*, if the law according to which all systems with the microstructure $<c_1, \ldots, c_n \mid o>$ have the property *E* can not be *deduced*, even in principle, from the laws that describe the behaviour and properties of the components $c_1, \ldots, c_n$ in isolation or in systems simpler than *S*.

However, Achim Stephan (1999, 37-44, see also Boogerd et al. forthcoming) has argued that there are in fact two quite different definitions of irreducibility implicit in Broad's writings, and that the failure to distinguish between these two has led the

[12] This was first pointed out by Pepper (1926,243): "It is only humanly speaking that anything is deducible. And what are strictly deducible are neither qualities nor events, but laws." Property reduction will be discussed in sections 6.2 and 6.3.

discussion of emergence astray. I believe this distinction is important, and I will try to explain it carefully here.

According to Broad, the supposedly emergent laws of chemistry and biology might in fact prove to be reducible: "Within the physical realm it always remains logically possible that the appearance of emergent laws is due to our imperfect knowledge of microscopic structure or to our mathematical incompetence." (1925, 81). However, the laws connecting phenomenal (or secondary) qualities with physical properties are *necessarily* irreducible. This is because phenomenal qualities are not *behaviourally analyzable*. The following quote from Broad's critique of Alexander's *Space, Time and Deity* sheds more light on this idea:

> [R]ed seems to me to be a quality of a certain motion-complex in one sense, and life to be a quality of a more elaborate complex in a very different sense. By saying that a body is living I just *mean* that its motions and other changes fit into each other and into the environment in certain characteristic ways. The statement is an analysis of its characteristic modes of change. But in saying that a motion is red I certainly do not *mean* that it is a vibration of such and such frequency. The statement is not an analysis of its characteristic mode of motion; but is the assertion that a property, which is not analysable in terms – such as velocity, frequency, etc., that apply directly to motions as such, occupies the same contour as a certain set of motions. (1921, 145)

So, roughly speaking, the behavioural unanalyzability of a property means that it can't be analyzed in terms of the behaviour (in the widest possible sense) of the related structures.

We can now understand better why Broad's archangel cannot deduce phenomenal qualities. The archangel has unlimited capabilities of calculation and he can directly perceive all microscopical structures, but he cannot know what is the smell of ammonia. He cannot know this, because it is not *behaviourally analyzable*, and therefore cannot be deduced from the behaviour of the related structures. Thus we have the first of the two definitions for irreducibility (Stephan 1999, 41):

> *Irreducibility*$_1$. Systemic properties that are not *behaviourally analyzable* are (necessarily) *irreducible*$_1$.

This corresponds nicely with the notion of irreducibility that the functional model of reduction implies. This model will be further discussed in section 6.3, but briefly speaking, central to this model is that a necessary condition for the reduction of a property is that it must be functionalized, that is, defined by its causal role. It seems clear that a property that cannot be analyzed in terms of behaviour cannot be functionalized, and on the other hand, a property that cannot be functionalized cannot be analyzed in terms of behaviour. Therefore, these notions would seem to be essentially the same. One significant difference is that the notion of functionalization is clearly defined, while the idea of behavioural analyzability remains rather vague.

Now I will turn to the other definition of irreducibility. According to Broad, even behaviourally analyzable systemic properties can be irreducible. This is the case when the behaviour of the system components over which a systemic property supervenes cannot be deduced from the behaviour of the components in other systems, even in principle. If the behaviour of the components cannot be deduced, neither can the properties that supervene on these components. (Stephan 1999, 42-43).

This other form of irreducibility is not explicitly discussed in *The Mind and its Place in Nature*. It comes up in other writings of Broad, but is never too precisely formulated. (Stephan 1999, 42-43). In the current debate on emergence and reduction, this kind of irreducibility is not often discussed. Stephan (1999, 43) has formulated it as follows:

> *Irreducibility*$_2$. A systemic property *E* of a system *S* is *irreducible*$_2$, if the behaviour of the components of *S* does not follow from the behaviour of these components in other systems simpler than *S*.

These two criteria of irreducibility are completely independent from each other. Even if the behaviour of the components of a system *S* cannot be deduced from the behaviour of these components in other systems, it is perfectly possible that the properties of *S* are behaviourally analyzable. On the other hand, even if the properties of a system *S* are not behaviourally analyzable, it is perfectly possible that the behaviour of the components of *S* can be deduced from the behaviour of these components in other systems. Both criteria are sufficient but not necessary conditions for the irreducibility of a systemic property. If we combine these two, we have the following definition of irreducibility (Stephan 1999, 43):

> *Irreducibility.* A systemic property is irreducible, if (i) it is not behaviourally analyzable, *or* (ii) if the specific behaviour of the system components, over which the systemic property supervenes, does not follow from the behaviour of these components in isolation or in simpler systems.

Thus we have two completely different forms of irreducibility. They also have completely different implications for the problem of downward causation (Stephan 1999, 44). The form that is based on the irreducibility of component behaviour (irreducibility$_2$) would seem to imply that there must exist a form of downward causation. If the behaviour of the components of a certain system does not follow from their behaviour in other systems, the only naturalistic explanation for this seems to be that the system or its properties somehow causally affect its components.

On the other hand, irreducibility$_1$ does not imply downward causation. It is perfectly possible that the behaviour of the components over which the behaviourally unanalyzable properties supervene follows from their behaviour in other systems. The real problem here is whether behaviourally unanalyzable properties can have any causal role at all. If not, they are epiphenomenal, and that is not a satisfying outcome. This is one of the main problems of emergence and the subject of Chapter 7.

3.6. Unpredictability

For Broad, the criterion for emergence is the synchronic concept irreducibility. In the evolutionarily oriented theories of Alexander and Lloyd Morgan, the criterion is the diachronic concept *unpredictability*. (Stephan 1999, 45). However, for the current discussion of emergence in the philosophy of mind, this concept is only marginally important, and in addition, it can be argued that the most significant form of unpredictability is in fact irreducibility. Therefore, in spite of its importance for the history of emergentism, I will discuss this characteristic only briefly.

It is important to note that what is at issue here is unpredictability *in principle*. The British Emergentists were not interested in our capabilities of prediction as human beings. They imagined idealized situations, where there were no limits to the cognizers' capabilities of calculation or to the information available to them. The

point was to show that even an archangel or a demon with unlimited resources could not predict emergent properties, because they are *in principle* unpredictable. (Stephan 1999, 46).

A property can be unpredictable in two ways (Stephan 1999, 47). (i.) A property can be unpredictable for the reason that the microstructure of the system that instantiates it is unpredictable before its first appearance. (ii.) A property can be unpredictable for the reason that it is *irreducible*. If a property is irreducible, it cannot be deduced even from a full knowledge of the microstructure, and thus it cannot be predicted before its first appearance, even when the microstructure is predictable. I have already discussed irreducibility in the last section, so here I will only go through the ways in which a property can be unpredictable without being irreducible.

First of all, a property of a system *S* is reducible but unpredictable if the behaviour of the components of *S* can be deduced from their behaviour in other systems, but some of the systems necessary for these deductions come into being only *after* system *S*. In this case, the appearance of this property cannot be even in principle predicted, even though the property is reducible. It is, however, questionable whether this kind of unpredictability has any interesting theoretical implications. (Stephan 1999, 54-55).

Another possibility is that there are reducible but unpredictable properties because the universe is fundamentally indeterministic, and the microstructures of the systems that instantiate these properties are for that reason unpredictable. However, the situation where a structure is unpredictable before its first appearance only because the universe is indeterministic has little to do with the problem of emergence. (Stephan 1999, 56-57).

The third and most interesting possibility is that a process that leads to the coming into being of a novel structure is *chaotic*, and thus even the reducible properties of this structure cannot be predicted. Here the problem is whether or not chaotic processes are predictable "in principle". If we imagine a Laplacian demon that knows the state of the universe at some earlier instants and all the laws of nature, can he predict chaotic processes? It seems that this creature could predict them only if he knew *all* the details of the earlier states of the universe *perfectly* – infinitely many decimals after the point, so to speak. Imagining such a creature comes close to absurdity. Therefore, it is relatively safe to say that chaotic processes can lead to structures that are in principle unpredictable. (Stephan 1999, 56-57).

In a short form and taking these possibilities into account, the characteristic of unpredictability could be formulated as follows (Stephan 1999, 68):

> A systemic property is in principle *unpredictable* before its first appearance (i.) if it is irreducible or (ii.) if the structure that instantiates it is unpredictable before its first appearance.

3.7. Downward causation

The last and perhaps most problematic characteristic of emergence theories is *downward causation*. At this point, I will only make some preliminary remarks. The debate on downward causation and its history will be discussed in section 7.2.

The idea of downward causation is explicit in the theory of Lloyd Morgan. For example, in one passage of *Emergent Evolution* he writes:

> The go of physico-chemical events at the level of life is not the same as that which obtains at the level of materiality only; the go of organic events at the level of effective consciousness is not the same as that which obtains at the level of vitality only. I speak of this alteration in the manner of go at any given level as "dependent on" the new and emergent kind of relatedness which there supervenes in the course of emergent evolution. So long as the words are used in purely naturalistic sense, one may say that the higher kinds of relatedness guide or control the go of lower-level events. (1923, 131).

As I already mentioned in chapter 3.5, Broad's irreducibility$_2$ also seems to imply a form of downward causation. If the behaviour of the components of a system cannot be deduced from their behaviour in other systems, something must be having an additional causal influence on them. Presumably also Lloyd Morgan had something similar in mind, although he uses ambiguous terms like "relatedness" and "involution" that make his ideas hard to understand. The way in which downward causation could be actually carried out is discussed neither by Broad nor Lloyd Morgan. In Alexander's system there is no room for downward causation, and it is rather explicitly denied (see for example 1920, ii, 12). (Stephan 1999, 58-64).

According to Stephan, downward causation can be interpreted in two ways: (i.) the *system* that has emergent properties causally influences the behaviour of its

components, or (ii.) the *emergent properties* themselves influence the behaviour of the components of the system. This of course depends on whether we assume that properties as such can have causal powers or just the systems that realize them. (Stephan 1999, 58).

The downward causation of Lloyd Morgan and Broad is most plausibly interpreted as exemplifying the first form, where the system has causal influence on its components. However, in the contemporary discussion of emergence, the second form, where the properties have causal influence on the components of the system, is more central. Here the problem is that physics seems to leave no place for these additional causal powers – the physical realm is causally closed. (Stephan 1999, 64-65).

Thus we have to face what Stephan (1999, 65) has dubbed the "Pepper-Kim-Dilemma" according to its first (Pepper) and most famous (Kim) advocators. The dilemma is this: (i.) If an emergentist wants to grant emergent properties a causal role, he must accept a form of downward causation and thereby deny the causal closure of the physical realm. (ii.) If an emergentist denies downward causation, he must accept that emergent properties have no causal role, so that they are epiphenomenal.

The problem of downward causation is closely related to the problem of mental causation, which is the problem of how mental properties can have a causal role in a world that is fundamentally physical. Often mental causation is seen as a form of downward causation. In spite of this, the two problems should be kept separate. Mental causation is possible without there being any general principle of downward causation, and downward causation is possible without mental causation. I believe this is sometimes forgotten in the emergence debate. In this work, the main focus is on the problem of mental causation, while the problem of downward causation is not so central.

4. Different Types of Theories of Emergence

In this chapter I will define three different types of theories of emergence: *weak, synchronic* and *diachronic*. The distinctions between these three will be made in terms of the characteristics introduced in the previous chapter. Again, I will follow closely Stephan's *Emergenz* (1999). There are, naturally, other ways of classifying theories of emergence. For example, O'Connor and Wong (2002) distinguish between *epistemological* and *ontological* emergentism, where the first is concerned with the limits of human knowledge on complex systems, and the latter is based on Broad's concept of emergence. I believe, however, that Stephan's classification is the most elaborate and accurate one, and I will not deal with the alternatives here.

4.1. Weak emergentism

Weak emergentism has the following defining characteristics (Stephan 1999, 66-67):

(i.) *Physical monism*. This is a form of *naturalism*: all entities in the world are composed of physical elements. Therefore, also emergent properties are instantiated by systems that are completely composed of physical elements.

(ii.) *Systemic properties*. There are systemic properties. A property of a system is systemic if none of the components of the system has it.

(iii.) *Synchronic determinism*. The properties and behavioural dispositions of a system are nomologically dependent on its microstructure. There can be no difference in the systemic properties without there being a difference in microstructure.

Weak emergentism is perfectly compatible with reductive physicalism. A reductive physicalist need not deny the existence of systemic properties, and physical monism and synchronic determinism should be accepted by all naturalistically oriented philosophers. (Stephan 1999, 67). Weak emergentism is the foundation for stronger forms of emergentism, but it is so weak that it is doubtful whether it deserves the name emergentism at all. A good example of a weak theory of emergence is the

theory of Mario Bunge (see section 2.5). Weak notions of emergence are also used in cognitive science in describing the properties of connectionist networks and in theories of self-organization[13].

4.2. Synchronic emergentism

Synchronic emergentism is weak emergentism supplemented with the characteristic *irreducibility* (Stephan 1999, 68). This is the form of emergentism that is based on the writings of C. D. Broad and that is most significant for the discussion of emergence in the philosophy of mind. As I pointed out in section 3.5, the notion of irreducibility can be divided into two parts:

> (iv.) *Irreducibility*. Systemic properties are irreducible, (a.) if they are not behaviourally analyzable, or (b.) if the behaviour of the components over which they supervene is irreducible. In both cases the systemic properties cannot be deduced from the behaviour and properties that the components show in isolation or in simpler systems.

From the unanalyzability of a property does not follow the irreducibility of the behaviour of the components over which it supervenes. And the other way around, from the irreducibility of the behaviour of the components does not follow the unanalyzability of a systemic property. Form (iv.a.) does not imply downward causation, while form (iv.b.) apparently does so. Therefore, we clearly have two different forms of synchronic emergentism. Neither form is compatible with reductive physicalism because of the thesis of irreducibility. (Stephan 1999, 68).

4.3. Diachronic emergentism

Diachronic emergentism is weak emergentism supplemented with the characteristics *novelty* and *unpredictability* (Stephan 1999, 69). In the history of

[13] For more on emergence in cognitive science, see Stephan (1999, 219-231), and in theories of self-organization, Stephan (1999, 232-246).

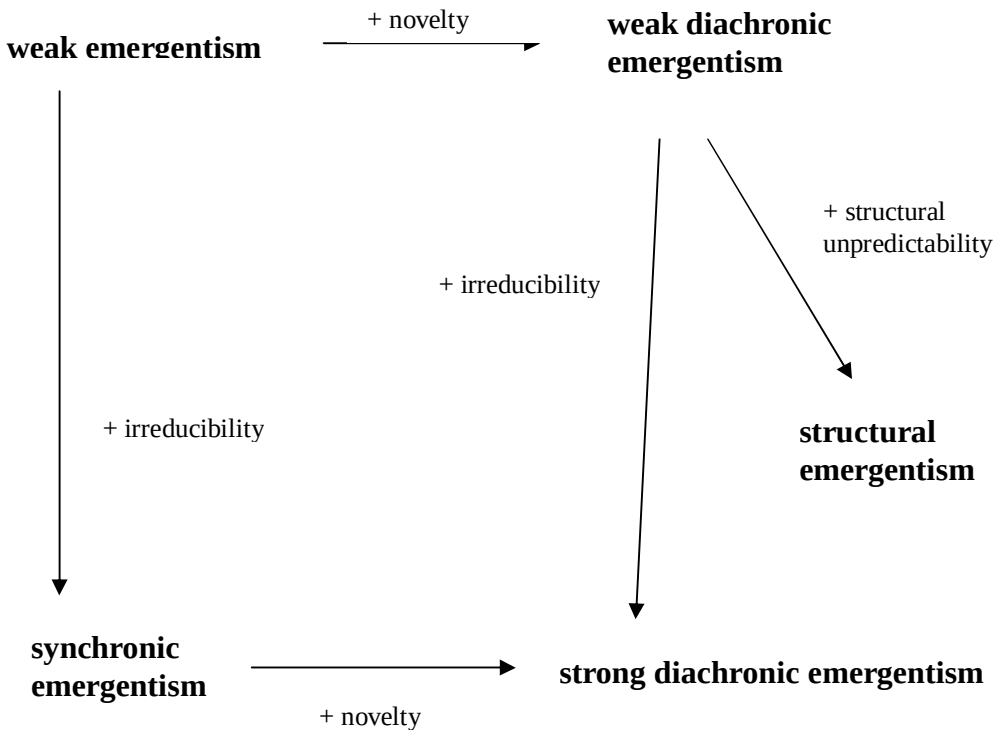
emergentism, this form is more prevailing than synchronic emergentism, but for philosophy of mind it is of little interest.

As was seen in section 3.6., a property can be unpredictable for two reasons: because the structure over which the property supervenes is unpredictable, or because the property is irreducible. As the early emergentists were convinced that the universe is deterministic and that all structures are in principle predictable, only the latter alternative is available for interpreting their theories. Therefore, the "classic" diachronic emergentism can be analyzed as weak emergentism supplemented with *novelty* and *irreducibility*. This diachronic emergentism can have two forms according to the two different forms of irreducibility. (Stephan 1999, 70).

The other type of diachronic emergentism, according to which there are structures that are in principle unpredictable, and therefore also the properties of these structures are unpredictable, could be called *structural emergentism*. On this form, the unpredictability of a property does not depend on its irreducibility. Although it is based on the limitations of the predictive powers of physical sciences, structural emergentism is compatible with reductive physicalism. (Stephan 1999, 70). It would be interesting to consider the theoretical possibilities of this kind of emergentism, but it has little to do with the problem of emergence in the philosophy of mind.

4.4. Summary of the Different Forms of Emergentism

The following picture illustrates the different forms of emergentism and the relations between them (based on Stephan 1999, 71):



Weak emergentism and weak diachronic emergentism are compatible with reductive physicalism. They are mainly applied in cognitive science, especially in theories of connectionist networks and self-organization. Strong diachronic emergentism differs from synchronic emergentism only with respect to the relatively unimportant characteristic of novelty. On the other hand, structural emergentism is completely independent of synchronic emergentism, and can have interesting applications, for example in evolutionary theories. It is in principle compatible with reductive physicalism. (Stephan 1999, 71-72).

The form that is most significant for contemporary philosophy of mind is *synchronic emergentism*. Because of the thesis of irreducibility, it is fundamentally incompatible with reductive physicalism. It can be divided into two forms in accordance with the two different forms of irreducibility. In the recent debate on emergence, mainly the first (iv.a.) form has been discussed. The rest of this work will be mainly concerned with this form of synchronic emergentism.

PART II

Emergence in the Philosophy of Mind

5. Physicalism

The return of emergentist ideas into philosophy mind in the 1970s was mainly due to the failure of reductive physicalism. The different reductionist theories were unable to credibly explain mental phenomena. This led to the development of various theories of non-reductive physicalism or property dualism. (Stephan 1999, 159). In this chapter, I will briefly go through the history of physicalism in the 20th century. In the end, I will also consider the question whether emergentism can be seen as a form of physicalism.

5.1. Semantic Physicalism and the Identity Theory

When physicalism was first formulated in the 1930s within the Vienna Circle, it was a theory about the language of science. According to this *semantic physicalism*, the physical language is the universal language of science, and all sentences of other sciences should be formulated in this language. Specifically, the talk of emotions and sensations in psychology is nothing but a shorter way of talking about observable behaviour (this is sometimes called *logical behaviourism*). All sentences of psychology should be translatable without loss of meaning into (long and complex) sentences of physical language. For example, the sentence "Cristina wants a beer" could be translated in the following manner: "If Cristina is asked whether she wants a beer, she will answer yes; if she is provided a beer, she will drink it; if she is in a bar, she will order a beer; etc." (Beckerman 1992a, 2-6; Churchland 1988, 23-25).

However, it soon became clear that this approach faces some serious problems. First of all, in the example above, the sentence is not fully translated, for at the end of the translation stands the seemingly innocent "etc.". If one tries to list all the behavioural dispositions associated with (for example) wanting something, the translation easily becomes indefinitely or infinitely long. Secondly, each of the partial sentences of the translation needs further limitations. For example, if asked whether she wants a beer, Cristina will answer yes *if* she doesn't want vodka even more; if she is in a bar, she will order a beer *if* she doesn't want to save her money, and so on.

Thirdly, the sentences of the translation are certainly not sentences of physical language, and translating them further into it would be difficult or impossible. (Beckerman 1992a, 2-6; Churchland 1988, 23-25).

These kinds of problems proved fatal to semantic physicalism. As a response to this failure, the *identity theory* (or *reductive materialism*) was developed in the 1950s. The starting point of this theory was a simple observation that was based on Frege's distinction between meaning and reference: even though a psychological term cannot be translated into physical language, it is possible that a psychological and a physical term refer to the same thing. So the main idea of the identity theory is strikingly simple: mental states *are* physical states of the brain. For example, the expressions "temperature" and "average molecuclar kinetic energy" may have different meanings and applications, but in fact temperature is identical to average molecular kinetic energy. In a similar way, each mental state or process is identical with a certain neurological state or process. It is the job of empirical research to reveal these identities. (Churchland 1988, 26-35; Beckermann 1992a, 6-11).

There are two main lines of argument aimed against the identity theory (Churchland 1988, 36-37; Beckermann 1992a, 8-11). The first is based on the *multiple realizability* of mental properties[14]. For example, the property of being in pain is quite differently realized in an octopus than in a human being. Even within the same human being the neurological realization of a property can change because of physical damage and the subsequent adaptation of the central nervous system. Thus it seems that there is no single type of physical state to which a certain mental state would always correspond. If this is so, there cannot be any universal identities between mental and physical states, and the identity theory cannot be true.

The other line of argument is based on Donald Davidson's (1970) *anomalous monism*. According to anomalous monism, even though all mental events are physical events (monism), the domain of the mental is not covered by strict laws (the anomality of the mental). It follows that there can be no strict laws connecting mental and physical properties and thus no laws expressing identities between mental and physical properties. The argumentation behind anomalous monism is complicated, and I will not go deeper into the details here.

[14] This line of argument goes back to Putnam (1967) and Fodor (1974).

These kinds of arguments lead to the fall of reductive materialism and to the development of different types of *non-reductive* theories. The idea was to hold on to physicalism but to deny that mental properties could be reduced to physical ones. The physicalism intended here was not *type* physicalism, according to which all mental types (or properties) are physical types (or properties), but rather *token* physicalism, according to which all events, and thus also mental events, are also physical events. (Kim 1998, 4-9; Fodor 1974).

5.2. Reduction

Before turning to theories of nonreductive physicalism, I will discuss the "classic" Nagelian model of reduction. Nagel discusses reduction in the 11th chapter of his classic work *The Structure of Science* (1961)[15]. For a long time, both the proponents and opponents of reductionism took this model for granted.

According to Nagel, reduction "is the explanation of a theory or a set of experimental laws established in one area of inquiry by a theory usually though not invariably formulated for some other domain" (p. 338). A theory T_1 is reducible to theory T_2 if the laws of T_1 can be deduced directly or with the aid of "bridge principles" (or "bridge laws") from the laws of T_2. Bridge principles are needed in most cases because the vocabularies of the theories are different and for that reason direct deductions are impossible. Bridge principles are usually thought of as nomological biconditionals that connect the terms of T_2 with corresponding terms of T_1. (Beckermann 1992b, 107).

Nagel's main example is the reduction of thermodynamics to statistical mechanics, particularly the derivation of the Boyle-Charles' law for ideal gases from the kinetic theory of gases (see Beckermann 1992b). As early as in the 19th century it could be shown that all the laws of thermodynamics can be deduced from statistical mechanics, and it was thought that for this reason thermodynamics had lost its autonomy as a science. In short, the Boyle-Charles's law for ideal gases,

$$pV = NkT ,$$

[15] Another classic exposition of this kind of reductionism is Oppenheim and Putnam (1958).

can be deduced in the following manner. First, the following law is deduced from statistical mechanics:

$$MV = (2/3)NE ,$$

where M is the average of the instantaneous momenta transferred from the molecules of the gas to the walls of the container and E the mean kinetic energy of these molecules. From this law, the Boyle-Charles's law can be deduced with the bridge laws:

$$p = M ,$$
$$(2/3)E = kT .$$

Therefore, the Boyle-Charles' law is a logical consequence of the principles of mechanics, when supplemented with appropriate bridge laws.

The Nagelian model of theory reduction played an important role in the philosophy of mind and philosophy of science in the latter half of the 20th century. However, later discussion has showed that this model is deeply inadequate. The defects of the Nagelian model and alternatives to it will be discussed in section 6.2.

5.3. Functionalism

Functionalism was for a long time and perhaps still is considered the principal form of nonreductive physicalism. The central idea of functionalism is that all mental states are defined by their *causal roles*, that is, the causal relations they bear to environmental effects on the body, other mental states and bodily behaviour. For example, the state of being in pain is characteristically caused by tissue damage and results in distress, annoyance, wincing and so on. Similarly, all mental states are be defined by the causal roles they play. (Churchland 1988, 36-42; Kim 1998, 4-9).

In functionalism, the relation of the mental to the physical is explained with the concept of *realization*. The idea is that mental properties are "realized" in physical properties, but are neither reducible nor identical to them. A mental property can be

realized in various different ways: for example, the state of being in pain is realized in different ways in different species. This is the idea of *multiple realization* already mentioned in section 5.1. It was also thought that if mental states are defined by their causal roles, they can be objects of science independently of their physical realization, and thus psychology and other sciences of the mental are *methodologically autonomous* from the physical sciences. (Churchland 1988, 36-42; Kim 1998, 4-9).

However, recent discussion has shown that functionalism may not be a nonreductive position after all. The functionalists, backed by the arguments of Fodor (1974) and Putnam (1967), thought that the multiple realizability of mental states guarantees the irreducibility of the mental[16]. However, the arguments for this were based on the Nagelian model of reduction, where bridge laws connecting mental and physical terms played a central role. Multiple realizability made the existence of bridge laws between the mental and the physical seem impossible. However, in newer models of reduction, bridge laws play no part at all. In the functional model of reduction that will be discussed in section 6.3, the functionalization of a property is a *necessary condition* for reduction. This means that on this model functionalism is a perfect example of a reductionist theory. (Kim 1998, 4-9; Stephan 1999, 165-173; Churchland 1988, 36-42).

5.4. Supervenience

The concept of *supervenience* was also supposed to explain the relation between the mental and the physical nonreductively. The concept originates from the writings of the moral philosophers G. E. Moore and R. M. Hare, who used it to describe the relation between natural properties and moral properties. Into philosophy of mind it was imported by Donald Davidson. In the article “Mental Events” (1970), where Davidson presents his anomalous monism, there is the following paragraph:

> Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events

[16] Putnam has later abandoned functionalism and strongly criticized it. See for example Putnam (1988).

> alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect. (1970, 214).

Here the term "supervenience" is used to describe a dependency relation that holds between the mental and the physical even though there are no laws connecting them. The idea of supervenience was quickly embraced by functionalists and other philosophers of mind as a nonreductive but ontologically physicalistic solution to the mind-body problem. (Beckerman 1992b, 94-95; Kim 1998, 4-9).

There are many variants of supervenience, but the form most relevant to the present discussion is "strong" supervenience that can be defined in two different ways (Kim 1997, 272-273):

1) Let *A* and *B* be sets of properties. *A* (strongly) supervenes on *B* iff, necessarily, for any *F* in *A*, if anything has *F*, there is a *G* in *B* such that the thing has *G*, and necessarily everything with *G* has *F*.

2) *A* (strongly) supervenes on *B* iff any two things (in the same or different possible worlds) indiscernible in *B*-properties are indiscernible in *A*-properties – that is, indiscernibility in respect of *B*-properties entails indiscernibility in *A*-properties.

Under certain assumptions these two definitions turn out to be equivalent, and they are often used interchangeably (Kim 1997, 273). I will follow this practice here.

When applied to mental and physical properties, supervenience means that there cannot be a change in the mental properties of a thing without there being a change in its physical properties, and that any two things that are indiscernible in all physical properties are indiscernible in all mental respects. In addition, for any mental property *M* there exists a physical property *P* that is necessarily sufficient for *M*. This means that if anything instantiates *M*, it also instantiates a certain physical property *P* such that *necessarily* anything with *P* has *M*. (Kim 1997, 273).

Supervenience is usually taken to be a relation of *dependence* or *determination*: if the mental supervenes on the physical, mental properties are dependent on physical properties or determined by them. This appears plausible, as the supervenience relation guarantees that when the physical properties of a thing are

fixed, its mental properties are also entirely fixed. However, the relation of dependence or determination is *asymmetric*: if *x* depends on, or is determined by, *y*, it cannot be that *y* also depends on or is determined by *x*. The relation of supervenience is *not* asymmetric: it is perfectly possible that *A* supervenes on *B* and *B* in turn supervenes on *A*. Thus, strictly speaking, supervenience does not involve dependence or determination. All that it states is a pattern of *covariance* between two sets of properties. This pattern can exist with or without any metaphysical determination or dependence between these properties. For example, it is conceivable that each set of properties depends on a third factor that creates the pattern of covariation. (Kim 1997, 273). However, in this work I will follow the usual practice and assume that supervenience is also a relation of dependence or determination.

In section 3.3, I described the layered model of nature. Central to this model is the idea that entities of a given level are completely composed of entities of lower levels. When supervenience is applied to this model, the natural result is *mereological supervenience*, according to which the properties of a whole are completely fixed by the properties and relations that characterize its parts. (Kim 1997, 278-279). For the discussion on emergence, this form of supervenience is the most relevant one.

It was thought for a long time that supervenience physicalism is a possible solution to the mind-body problem. However, later discussion has showed that supervenience alone is not enough to constitute a mind-body theory. The most obvious reason for this is that supervenience is consistent with various classic positions on the mind-body problem that are completely different from each other and even conflicting. For example, type physicalism implies supervenience, but so does synchronic emergentism. The first theory is fundamentally reductionist, the second by definition non-reductionist. Supervenience is even compatible with dualistic theories like parallelism. Supervenience is a shared commitment of various different mind-body theories, and not itself a position one can take on this issue. (Kim 1997, 274-275).

5.5. Minimal physicalism

Here at the end of this chapter I will briefly discuss the nature of physicalism, or what is required of a theory for it to be called physicalistic. Specifically, I will

consider whether there are grounds for calling emergentism a physicalistic theory. This question is particularly interesting because emergentism is often seen as a form of nonreductive physicalism.

One possible answer is that *token physicalism* is enough to capture the idea of physicalism. According to token physicalism, every particular thing (object, process, event etc.) in the world is a physical particular. Therefore, all mental particulars are physical particulars. At first glance this might seem like a good answer, but in fact token physicalism is an extremely weak form of physicalism. It is compatible with the idea that there are no law-like connections between mental and physical properties. It is also consistent with a form of dualism, namely property dualism, as it does not rule out the possibility that some particulars have non-supervenient mental properties. Therefore, token physicalism alone is too weak to constitute minimal physicalism. (Stoljar 2001).

Another possible answer is that *supervenience* is enough to make a theory physicalistic (Stoljar 2001). Supervenience excludes the possibility that mental properties "float freely": if the physical nature of a thing is fixed, its mental nature is completely fixed too. Jaegwon Kim has repeatedly claimed that supervenience defines "minimal physicalism" (see for example 1998, 14-15). However, what he means by this is that supervenience is a shared commitment of various different physicalistic positions. As was seen in the last section, supervenience is merely the statement of a pattern of covariation. It is compatible with various mind-body theories, some of which are certainly non-physicalistic (for example dualistic epiphenomenalism). Therefore, supervenience alone can not be enough to guarantee physicalism.

However, perhaps supervenience is enough if it is supplemented with an appropriate principle of *physical realization*. One way to formulate such a principle is this (Kim 1996, 74):

> If a system *x* has at a time *t* a mental property *M* (or *x* is at *t* in a mental state *M*), then *x* is a material thing, and *x* has *M* in virtue of the fact that *x* has at *t* a physical property *P* that realizes *M* in *x* at *t*.

This principle states that systems have mental properties in virtue of the physical properties that realize them. However, the principle is purely ontological: it does not say *how* the mental properties are realized by physical properties. Some

philosophers have argued that ontological claims like this are not enough to capture the idea of physicalism, what is also needed is the requirement that everything can be physically explained.

The idea that ontological supervenience is not enough for physicalism has been elaborated by Terence Horgan in his article "From Supervenience to Superdupervenience" (1993). According to Horgan, "the sort of inter-level relation that would confer materialistic 'respectability' on higher-order properties and facts would be not bare ontological supervenience, but superdupervenience – ontological supervenience that is robustly explainable in a materialistically acceptable way" (p. 577). The point is that a genuine physicalistic position cannot include unexplained supervenience relations, only those whose existence is materially explainable.

There are stronger forms of physical realization that "robustly explain" the supervenience relation. Ansgar Beckermann (1997, 310) has proposed the following formulation that is based on the ideas of C. D. Broad:

> A property *F* is realized within a system *S* through the property *G* if and only if *S* has the property *G* and it follows on the basis of laws of nature which generally apply to objects with the property *G*, that *S* if it has the property *G*, possesses all features which are characteristic of *F*, or that *G* in the system *S* possesses all features which are characteristic of *F*.

This definition does not allow any "explanatory gaps"[17] between mental and physical properties: if a mental property is physically realized, it is possible to deduce all its characteristic features from knowledge of the realizing property and the laws of nature. That is, if a mental property is physically realized, it is also reducible.

Jaegwon Kim (1997) has also put forward a model of physical realization that materialistically explains supervenience. On this model, mental properties are construed as functional second-order properties of physical properties. Second-order properties are defined as follows (1997, 280):

> *F* is a *second-order property* over set *B* iff *F* is the property of having some property *P* in *B* such that *D*(*P*), where *D* is a condition on members of *B*.

[17] The "explanatory gap" (Levine 1983) will be further discussed in section 6.1.

Functional second-order properties are the ones whose specification *D* involves causal relations. The physical properties that fill the specification *D* *physically realize* the property *F*. Let us take for example the property of being in pain. In this case, the set *B* consists of physical properties, and *D* specifies a certain causal role characteristic of being in pain – i.e., that it is caused by tissue damage and results in distress, wincing etc. To be in pain is to have a physical property that fills the causal role specified by D. Let us assume that for human beings this property is P. Then for human beings, being in pain is nothing over and above having property P. (Kim 1997, 279-286. More on this subject in section 6.3.)

With these stronger principles of physical realization, *emergent properties no longer count as physically realized properties*. According to synchronic emergentism, emergent properties are *irreducible*. At this point, we may assume that this means that they are functionally or behaviourally unanalyzable. Beckermann's definition requires that *all* the characteristics of the realized property must be deducible from knowledge of the realizing property and the laws of nature. If emergent properties are functionally or behaviourally unanalyzable, this is not possible. Kim's definition requires that physically realized properties must be functionally definable, and this is of course impossible for properties that are functionally unanalyzable. If physicalism requires that all properties must be physically realized in these stronger senses, *emergentism is not a form of physicalism.* (Stephan 1999, 174-178).

To sum up: Token physicalism and supervenience are certainly not enough to guarantee physicalism. Supervenience with the weaker form of physical realization may be enough for some, but it leaves open the possibility that there are physically realized properties that can never be physically explained. The stronger forms of physical realization rule out the possibility that emergent properties are physically realized. There are, therefore, good grounds for arguing that emergentism is not a physicalistic theory after all. It might be more appropriate to see it as a form of *property dualism*.

This would have some interesting implications. At the moment, emergentism is arguable the only viable form of nonreductive physicalism. If it is not a physicalistic position after all, there seems to be nothing left for a nonreductive physicalist. He must either abandon physicalism and face the problems of property dualism, or

embrace physicalism and deny that there are irreducible properties[18]. Or develop a new kind of theory.

[18] Kim has been defending a similar view for some time now, see for example 1993 or 1998.

6. Qualia Emergentism

6.1. Arguments for Qualia Emergentism

In this section I will go through certain important arguments that can be seen as arguments for qualia emergence, although the writers do not refer to emergentism or use the concept emergence. They were put forward before the actual emergence debate took off, and in some ways form the basis for it. The arguments that I will discuss are Frank Jackson's knowledge argument and Joseph Levine's explanatory gap argument. I will also compare Jackson's argument with Broad's archangel example. The whole section is largely based on Stephan (1999, 185-196).

Frank Jackson's knowledge argument is perhaps most clearly formulated in his article "What Mary Didn't Know" (1986). The argument introduces the nowadays philosophically famous scientist Mary:

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in *completed* physics, chemistry and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. If physicalism is true, she knows all there is to know. For to suppose otherwise is to suppose that there is more to know than every physical fact, and that is just what physicalism denies.
>
> […]
>
> It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. This is rightly described as *learning* – she will not say "ho, hum". Hence, physicalism is false. (Jackson 1986, 291).

The point is that Mary knows all the physical facts there are to know, but she still doesn't know what it is like to see red, for example. Therefore, physicalism cannot be true.

The "classic" answer to the knowledge argument is that Mary doesn't learn anything new in the sense of propositional knowledge, she merely acquires new

abilities, for example the ability to imagine red things at will[19]. Therefore, the argument fails to show that reductive physicalism is false. The first to propose this answer was Laurence Nemirow:

> [S]ome modes of understanding consist, not in the grasping of facts, but in the acquisition of abilities – for example, understanding a language, or understanding a rule (in Wittgenstein's sense). As for understanding an experience, we may construe that as an ability to place oneself, at will, in a state representative of the experience. I understand the experience of seeing red if I can at will visualize red. […] We can, then, come to terms with the subjectivity of our understanding of experience without positing subjective facts as the objects of our understanding. (Nemirow 1980, 475-476).

There are several convincing arguments against this answer. For example, we can imagine a person with no abilities of visual imagination. He can't imagine what it is like to see red. However, at the moment when he sees something red, he knows what it is like to see red, even though he does not have the ability "to place oneself, at will, in a state representative of the experience". Therefore, the ability to imagine red things at will is neither a necessary nor a sufficient condition for knowing what it is like to see red. (Gertler 1999, 320).

Another answer to Jackson's argument is that Mary doesn't really learn any new facts, she merely learns to represent facts she already knows in a new way. The most prominent proponent of this answer has been Paul Churchland:

> In short, the difference between a person who knows all about the visual cortex but has never enjoyed a sensation of red, and a person who knows no neuroscience but knows well the sensation of red, may reside not in *what* is respectively known by each (brain states by the former, qualia by the latter), but rather in the different *type* of knowledge each has *of exactly the same thing*. (Churchland 1985, 24).

The most obvious problem with this answer is that the knowledge of a certain brain state and the knowledge of the phenomenal quality of seeing red appear to be so fundamentally different that it seems impossible that they could be knowledge of one and the same thing. C. D. Broad held a similar view on the matter and it is further backed up by Levine's explanatory gap argument (see below).

[19] This answer was first put forward by Laurence Nemirow (1980) against Thomas Nagel's similar argument ("What Is It Like to Be a Bat", 1974). See also Lewis (1988) for a careful analysis of Jackson's argument and a similar answer.

Martine Nida-Rümelin (1995) has also thoroughly analyzed the argument and quite convincingly argued that Mary gains new factual knowledge after her release. Nida-Rümelin's argumentation is very subtle and I cannot go very deep into it here, but I will try to summarize the main points.

Nida-Rümelin distinguishes between phenomenal and non-phenomenal knowledge[20]. Mary has all the non-phenomenal knowledge of mental states before her release, she knows for example that the colour of the sky appears blue to normally sighted people. However, "[t]here is a kind of factual knowledge about phenomenal states that is accessible only to an epistemic subject who knows the kind of phenomenal state at issue by personal experience" (1995, 237). Mary cannot possibly have this phenomenal knowledge before her release.

Nida-Rümelin leaves open the ontological question of whether this means that there are non-physical facts. At least the argument shows that all descriptions of conscious beings given in purely physicalistic terminology are necessarily incomplete. Even if this is all the argument shows, it is still a strong argument for the irreducibility of qualia: if all the characteristic features of qualia cannot be expressed in physicalistic terminology, it is hard to see how qualia could be reduced to physics.

It is obvious that Jackson's argument resembles in many ways C. D. Broad's archangel example (see section 2.3). However, there is one important difference, and because of this difference, Broad's thought experiment manages to avoid many of the problems associated with Jackson's argument. For example, the arguments against Jackson's argument mentioned above do not apply to the archangel example. The central ideas of Broad's argument can be summarized as follows (based on Stephan 1999, 186):

1. The mathematical archangel knows the exact atomic structure of all chemical compounds and the exact neuronal structure of our sense organs and our central nervous system, and he knows all the relevant laws of nature.

[20] Earlier David Lewis (1988) has argued that Jackson's argument leads to the Hypothesis of Phenomenal Information (HPI). This hypothesis states that in addition to physical information there is a different kind of information to be had, namely phenomenal information. Lewis considers this to be an unacceptable outcome, mainly because he sees it as violating physical causal closure.

2. The mathematical archangel cannot, even with all his physical and physiological knowledge, deduce all the characteristic features of phenomenal states (for example, of the sensation of smelling sulphur).

Therefore, even if the phenomenal states are determined by neuronal events, they are explanatorily irreducible and thus necessarily emergent.

The difference between the two arguments is that Broad's archangel doesn't even know that there are qualitative experiences associated to neurological states, because he cannot deduce this from physical and physiological facts alone. In Nida-Rümelin's terminology, he doesn't even have the non-phenomenal knowledge of which phenomenal states are associated with which neurological states. Therefore, it can't be argued that the archangel is merely lacking certain abilities or ways of representing things: there are clearly facts about the world that he doesn't know. The crucial point is then whether are enough grounds for holding the second premise true.

In section 3.5, I have already discussed some of the grounds for this. In addition, Broad has argued that it would be a caterogical mistake to reduce phenomenal properties to physical ones: "Let us suppose, for the sake of argument, that whenever it is true to say that I have a sensation of a red patch it is also true to say that a molecular movement of a certain kind is going on in a certain part of my brain. There is one sense in which it is plainly nonsensical to attempt to reduce the one to the other." (1925, 622). The reason why it is nonsensical is that here are some questions that can be asked of the other characteristic which are nonsensical to raise about the other characteristic. "About a molecular movement it is perfectly reasonable to raise the question: 'Is it swift or slow, straight or circular, and so on?' About the awareness of a red patch it is nonsensical to ask whether it is a swift or a slow awareness, a straight or a circular awareness, and so on." (1925, 622-623).

These considerations are somewhat similar to Joseph Levine's (1983, 1993) explanatory gap argument. While Broad tried to show that it is caterogical failure to identify conscious sensations with neurological events, the point of Levine's argument is to show that psycho-physical identity statements are not fully explanatory: they always leave a significant *explanatory gap*. We can compare for example sentences like:

1) Pain is the firing of C-fibers.
2) Heat is the motion of molecules.

According to Levine, the difference between these two statements is that (2) expresses an identity that is fully explanatory, with nothing crucial left out, while sentence (1) does leave something crucial unexplained. Sentence (2) is explanatory in the sense that our knowledge of chemistry and physics makes it understandable how motion of molecules can play the causal role associated with heat. In addition, the causal functions associated with heat exhaust the notion of heat – once we understand them, there is nothing more to understand. (1983, 357)

The situation with sentence (1) is different. Certainly part of the concept of pain is that it is a state that plays a certain causal role: it is usually caused by tissue damage and results in wincing, annoyance etc. Sentence (1) explains how these causal functions are carried out. However, "there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is *why pain should feel the way it does*!" (Levine 1983, 357). There seems to be nothing about C-fiber firing that makes it fit better with the phenomenal properties of pain than with any other phenomenal properties. The identification of pain with C-fiber firing makes the way pain feels into merely a brute fact. (1983, 357).

The following paragraph wonderfully summarizes the central points of Levine's argument:

> What seems to be responsible for the explanatory gap, then, is the fact that our concepts of qualitative character do not represent, at least in terms of their psychological contents, causal roles. Reduction is explanatory when by reducing an object or property we reveal the mechanisms by which the causal role constitutive of that object or property is realized. Moreover, this seems to be the only way that a reduction could be explanatory. Thus, to the extent that there is an element in our concept of qualitative character that is not captured by features of its causal role, to that extent it will escape the explanatory net of physicalistic reduction. (Levine 1993, 134)

The ideas behind explanatory reduction will be further discussed in section 6.3. Before that, I will briefly discuss some general problems of reduction.

6.2. Problems of Reduction

Nagel's derivational model of reduction described in section 5.2 dominated philosophical discussions of reduction in the latter half of the 20th century. Central to this model are "bridge laws" that connect terms of the theory targeted for reduction with terms of the base theory. However, later discussion has showed that this model is in many ways deficient. There are very few if any examples of pure Nagelian reductions – even paradigm cases like the reduction of thermodynamics to statistical mechanics have turned out to be much more complex and debatable than Nagel and others thought. (Dupré 2000). The Nagelian model does not adequately describe the actual process of reduction in the sciences, and most importantly, it does not capture the central ideas behind reduction (see the next section).

Arguably the best alternative to Nagel's model at the moment is the functional model of reduction that will be discussed in the next section. Another noteworthy alternative is the model developed by C. A. Hooker (1981). According this model, if a theory T_1 is to be reduced to theory T_2, it is not necessary to deduce the laws of T_1 from the laws of T_2. It is enough to deduce from T_2 an *image* of T_1. In other words, it is essential to show that there are properties expressible in terms of T_2 which play (almost) the same role as the properties definable in terms of T_1. For example, the law $MV = (2/3)NE$ of statistical mechanics is an image of Boyle-Charles's law in thermodynamics. Similarly, images of all other laws of thermodynamics can be deduced from statistical mechanics. (Beckermann 1992b, 108-109)

On these alternative models of reduction, bridge laws play no part at all. This is interesting, because for a long time the discussion of reduction in philosophy of mind was centred round bridge laws. It was thought that the mere existence of bridge laws between mental and physical properties was enough to guarantee the reducibility of the mental. The central question was whether or not it is possible to connect all mental properties with physical properties with the help of bridge laws. Without the Nagelian model of reduction as a background, this whole question has become irrelevant. In fact, the existence of bridge laws between the mental and the physical is perfectly compatible with emergentism, epiphenomenalism and other nonreductive theories (Kim 1999, 12-13). If bridge laws are not further explained, they tell us nothing about the relation between the mental and the physical.

Another point is that the Nagelian account of reduction focuses on theories, while in the philosophy of mind the main interest has been in the reduction of properties. I will now turn to Robert Cummins' (1983) analysis of property explanation, as it neatly clarifies some issues related to emergence and reduction.

Cummins distinguishes between two kinds of theories: *transition theories* that are designed to explain changes and *property theories* that are designed to explain properties. It is widely agreed that the only way transition theories can explain changes is by *causal subsumption* – that is, by subsuming events or types of events under causal laws. Cummins argues that the main focus in philosophy of science has traditionally been on transition theories and that scientific methodology has often been identified (tacitly or explicitly) with the methodology of causal subsumption. On the other hand, the methodology of the analytic strategy of explanation involved in property theories has been largely forgotten, although the analytic strategy is as old as atomism. (Pp. 1-14).

According to Cummins, "[m]any of the most pressing and puzzling scientific questions are questions about properties, not about changes. […] Good property theories are wonderfully satisfying: we know how temperature is instantiated, how inheritance is instantiated, how electricity is instantiated, how solubility is instantiated." (P. 15). The characteristic question answered by a property theory is: What is it for system *S* to have property *P*? The natural way to answer this question is to give an analysis of *S* that explains *S*'s having *P* by appeal to the properties of *S*'s components and their mode of organization. (P. 15)

This kind of property explanation can also be called *microreduction* (Beckermann 1992b, 110). According to Cummins, property explanations can be represented in the following scheme (p. 17):

(i) Anything having components $C_1 \ldots C_n$ organized in manner O – i. e., having analysis $[C_1 \ldots C_n, O]$ – has property P;
(ii) S has analysis $[C_1 \ldots C_n, O]$;

(iii) S has property P.

Cummins calls laws like (i) "instantiation laws" and laws like (ii) "composition laws". However, this is not yet a sufficient explanation of the property: "The instantiation laws […] obviously call for explanation themselves." (p.18). It is

easy to see how this comes close to the ideas of C. D. Broad. According to Broad, there are laws like (i) even for emergent properties and they call for explanation, but the point is that they *cannot* be further explained, even in principle.

The point of all this is that models of reduction in which the mere correlation of properties suffices for reduction are too weak to provide an adequate framework for the discussion of emergence. What the early emergentists were concerned with and what is needed in the current debate is a model of *explanatory reduction* of properties. In the next section, I will describe one such a model that is based on the functionalization of properties. On this model, it is easy to see what makes certain types of mental properties fundamentally irreducible and thus emergent.

6.3. The Functional Model of Reduction

Jaegwon Kim, one of the most influential philosophers of mind, was earlier convinced that Nagelian bridge principles are sufficient for reduction. Later he has admitted that the Nagelian model is deeply inadequate and that the whole discussion was largely beside the point (see for example 1999, 12-13). Nowadays Kim is perhaps the most prominent defender of the new functional model of reduction. However, as Kim himself has pointed out (1999, endnote 15), the fundamental ideas of this model can be found in the writings of Lewis (1966) and Armstrong (1968). Presumably the first explicit formulation of the model is in Joseph Levine's article "On Leaving Out What It Is Like" (1993).

According to Levine, the basic idea of reduction is that it should explain what is being reduced. This is accomplished when we can see why, given the facts cited in the reduction, the reduced thing behaves or appears to be as it does. For example, the chemical properties of H_2O explain why water has the superficial macroproperties that we can perceive. And the reason for this is that H_2O is *causally responsible* for the properties of water: we can explain the transparency of water, its boiling and freezing points etc. by appealing to the causal properties of H_2O. In order to see this, we need to give a causal definition for the macroproperties to be reduced. (Levine 1993, 131-132). So we get the following picture:

> Our concepts of substances and properties like water and liquidity can be thought of as representations of nodes in a network of causal relations, each node itself capable of further reduction to yet another network, until we get down to the fundamental causal determinants of nature. We get bottom-up necessity, and thereby explanatory force, from the identification of the macroproperties with the microproperties because the network of causal relations constitutive of the micro level realizes the network of causal relations constitutive of the macro level. Any concept that can be analysed in this way will yield to explanatory reduction. (Levine 1993, 132).

This kind of explanatory reduction is a two-stage process:

> Stage 1 involves the (relatively? quasi?) a priori process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are. (Levine 1993, 132).

Jaegwon Kim's functional model of reduction is essentially the same, but much more elaborately described. Kim takes as a starting point the reduction of the gene to the DNA molecule. To reduce the gene, we must first prime it by giving it a *functional* interpretation, that is, we need to construe it in terms of its causal roles. The property of being a gene is the property of having a certain property that fills a certain causal role – namely, transmitting information of phenotypic characteristics from parents to offsprings. In this world, the DNA fills this causal role, and we have theories that explain exactly how the DNA molecule does this. With all this, we can say with good reason that the gene has been reduced to DNA. (Kim 1999, 10).

In general, the reduction of a property *E* proceeds as follows. First, *E* must be functionalized. This is accomplished by defining *E* as a functional second-order property. Second-order properties can be characterized as follows (Kim 1998, 20):

> *F* is a second-order property over set *B* of base (or first-order) properties iff *F* is the property of having some property *P* in *B* such that *D*(*P*), where *D* specifies a condition on members of *B*.

Second-order properties are second-order in the sense that they are generated by existential quantification over first-order properties. The first-order properties are the *realizers* of the second-order property. For example, if the base set B comprises of

colours, the property of having a primary colour can be thought of as a second-order property: having a primary colour is having a property *P* in *B* such that *P* = red or *P* = green or *P* = blue. In this case, being red, being green and being blue are the realizers of having a primary colour. If the realizers are physical, the second-order property is *physically realized.* It is clear that multiple realizers of a property are allowed, so *multiple realizability* is easily accounted for in this model. (Kim 1998, 20; 1999, 9-13)

Functional properties are the ones whose specification *D* involves causal relations. That is, functional properties are second-order properties defined in terms of causal relations among first-order properties. For example, dormitivity is a property that a substance has if it causes people to sleep. Various drugs have this property, but in virtue of different chemical realizers. Therefore, dormitivity is a second-order functional property of chemical substances. (Kim 1998, 20-21).

After the property *E* has been functionalized, the next step in the reduction is to find the realizers of *E* and the theories that explain how the realizers are able to fill the causal roles constitutive of *E*. This is of course a task of scientific research. (Kim 1999, 11-12).

It is easy to see that this model differs radically from the Nagelian model of reduction. Most importantly, in the functional model, bridge laws play no part at all. The model also has some interesting implications for functionalism (see section 5.3). According to functionalism, mental states are defined by their causal roles. Because these causal roles can be filled with various different physical properties, mental states cannot be Nagel-reduced to physical properties. However, according to the new model of reduction, the functionalization of a property is a *necessary condition* for its reduction. This means that if the functionalist conception of the mental is correct, reduction of the mental is at least in principle possible. This is interesting, because earlier functionalism was thought to be the most important form of nonreductive physicalism. (Kim 1998, 97-103).

The functional model also makes possible the explanation of Cummins' instantiation laws (see the previous section). The problem was to explain laws like "any system *S* with the microstructure *O* has property *P*". The explanation is this: having *P* is having a property with a causal role *C*, and system *S* has a (microstructural) property *Q*, which fills causal role *C*. In systems like *S*, having *P* *consists in* having *Q*, or in other words, having *P* is *nothing over and above* having *Q*. (Kim 1998, 103-118)

This makes the functional model truly a model of *reduction*: central to the concept of reduction is that what has been reduced no longer needs to be regarded as an independent existent. Therefore, it is a little bit misleading to talk of second-order *properties* as objects of reduction. These "properties" are merely results of existential quantification over first-order properties, and it would perhaps be more appropriate to talk of second-order *concepts* or *descriptions*. (Kim 1998, 103-118; 1999, 13-16).

The functional model of reduction supports the central ideas of emergentism. Emergent properties are emergent because they are irreducible, and on this model, irreducible properties are properties that cannot be functionalized. The chemical and biological properties that the early emergenists thought to be emergent can probably be functionalized, but there are good reasons to believe that certain mental properties – that is, phenomenal properties or qualia – can never be functionalized. Then we face the already familiar dilemma: either these properties have no causal powers and they are mere epiphenomena, or they have causal powers and the causal closure of the physical is breached. This dilemma will be further discussed in Chapter 7.

6.4. The New Qualia Emergentism

I will now sum up the discussion so far by giving the central argument of the *new qualia emergentism* (based on Stephan 1999, 195):

1. Reduction is explanatory only when it reveals the causal mechanisms that fill the causal role of the reduced property, that is, only if it follows the functional model of reduction.
2. The essential characteristics of qualia are not captured by their causal roles.
3. Properties that cannot by explanatorily reduced are emergent properties.

Therefore, qualia are emergent properties.

If one wants to avoid an emergentist position and deny the validity of this argument, there are basically two alternatives. The first is to attack the first premise

and somehow show that explanatory reduction is possible without the functional model of reduction. This is of course possible, and probably the best way to attack the argument. To my knowledge, at this point the only model of explanatory reduction we have is the functional model, but others can be developed. The second alternative is to show that qualia can be adequately defined by their causal roles. This seems to me impossible for the several reasons already discussed in various parts of this work.

If one accepts this argument and assumes an emergentist position, one has to face the problem of mental causation that will be discussed in the next chapter. At this point it can already be said that there is no solution to this problem in sight. For if phenomenal properties cannot be defined by their causal roles, how can they play a causal role? And if they have no causal powers, in what sense do they exist?

One important question is: what kind of psycho-physical position does the argument entail? Some writers like Levine (1983, 1993) claim that the premises (1) and (2) do not have any ontological consequences: they do not show that qualia are non-physical phenomena, they merely throw the burden back to the physicalists to show why we should think that they are physical. Indeed, the argument by itself does not show that qualia are non-physical, all it shows is that they are not explanatorily reducible. However, as we have already seen in section 5.5, there are good grounds for arguing that properties that cannot be explanatorily reduced do not fit into a physicalist framework. If physicalism requires that mental properties are physically realized in a sense that does not allow any explanatory gaps between the mental and the physical, then the new qualia emergentism is not compatible with physicalism.

7. The Problem of Mental Causation

Now I will finally turn to the problem of mental causation that has been mentioned so many times in this work. First of all, it must be mentioned that it is not a problem just for emergentism: it has been one of the main subjects in the philosophy of mind over the last decades, and broadly speaking, it has been a central theme in philosophy at least since Descartes. Another thing to keep in mind is that there is no single problem of mental causation. Jaegwon Kim (1998) has distinguished between three separate problems of mental causation: the problem of mental anomalism, the problem of content externalism and the problem of causal exclusion. Only the last of these is directly relevant for the current theme, but it is also the most fundamental one. The problem is to answer this question: How is it possible for the mind to exercise its causal powers in a world that is fundamentally physical? Or in other words: Given that every physical event that has a cause has a physical cause, how can a mental cause also be possible? (Kim 1998, 32-38).

The significance of this problem can hardly be overestimated. Jerry Fodor has aptly characterized the situation as follows:

> I'm not really convinced that it matters very much whether the mental is the physical; still less that it matters very much whether we can prove it is. Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying, … if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world. (1989, 77; quoted in Kim 1992, 135).

Mental causation is often seen as a form of "downward causation", where the properties of a system exert causal influence "downward" on the components of the system. The term comes from Donald Campbell (1974), and the idea has been developed especially by the famous neurobiologist Roger Sperry. The concept has also played a central role in the recent debate on emergence. However, it is important to keep the problem of mental causation separate from the problem of downward causation. Mental causation is possible without downward causation, and downward causation is possible without mental causation. The problem of downward causation will be further discussed in section 7.2.

As I already mentioned in section 3.7, Achim Stephan has formulated the problem emergentism has to face as a dilemma. The emergentist must either (i) accept ephiphenomenalism and deny that mental properties can play a causal role in this world, or (ii) claim that mental properties have causal powers and face the problem of mental causation. Stephan calls this the "Pepper-Kim-Dilemma of emergentism", according to its first and most prominent proponents. Stephen Pepper has presented the dilemma as early as 1926, although he did not consider option (ii) to be a real alternative. Most clearly and thoroughly the dilemma has been discussed by Jaegwon Kim, who is also one of the most important critics of mental causation. In the next section, I will briefly discuss Pepper's argument and the reasons why epiphenomenalism is considered such a bad alternative. In section 7.3, I will discuss the second horn of the dilemma, the problem of causal exclusion. There I will present Kim's supervenience argument against mental causation. In the last section I will try to see if there is any way out of it.

7.1. Epiphenomenalism

According to Pepper (1926), theories of emergence are theories of change. There are three kinds of changes in the world: (i) chance occurences or cosmic irregularities, (ii) "shifts", changes where one characteristic replaces another and that can be described as invariable succession, for example changes of position, and (iii) emergence, which is a cumulative change, "a change in which certain characteristics supervene upon other characteristics, these characteristics being adequate to explain the occurrence on their level" (p. 241). In the discussion of emergence, the possibility of cosmic irregularities is not at issue, the question is whether there can be emergent changes in addition to shifts.

Pepper does not even consider the possibility of downward causation. Apparently he assumes some sort of Laplacian determinism at the physical level, and this of course makes downward causation impossible. Therefore, according to Pepper, emergentism is subject to the following dilemma: either emergent changes are not really cumulative or they are epiphenomenal. An emergentist must accept that emergent changes are really just shifts or face epiphenomenalism, which Pepper considers a "metaphysically unsatisfactory" alternative (p. 241).

Pepper's argumentation for his dilemma is not very convincing. The problem is that he assumes that emergent qualities or properties are obviously epiphenomenal and concentrates on emergent laws. However, it is not at all self-evident that emergent qualities are epiphenomenal – many of the early and current emergentists have thought the contrary. On the other hand, the early emergentists were more concerned with emergent properties than laws, and the irreducible laws they discussed were property laws, not "laws of change". Therefore, Pepper's critique is not really a challenge for emergentism. However, it can be seen as a beginning for a line of argumentation that threatens the plausibility of emergentism more than anything else. Pepper failed to show that an emergentist must accept epiphenomenalism, but Kim's arguments against mental causation seem to accomplish this.

But why is epiphenomenalism such a bad alternative after all? The reason for this is most aptly formulated by Samuel Alexander in this often quoted passage:

> [Epiphenomenalism] supposes something to exist in nature which has nothing to do, no purpose to serve, a species of *noblesse* which depends on the work of its inferiors, but is kept for show and might as well, as undoubtedly would in time, be abolished. (1920, ii, p. 8).

What Alexander is emphasizing here is that *to be real is to possess causal powers*. Kim calls this "Alexander's Dictum", and it is an important part of Kim's own philosophy. Kim's ontological assumptions rule out the possibility of epiphenomenal properties. According to Kim, properties "cut reality at its joints", that is, they must be involved in laws and causal relations in an appropriate way. If a property is not involved in laws and causal relations, then it is not a genuine property. (See Loewer 2001). It is obvious that with these ontological commitments, epiphenomenal properties cannot be real properties.

However, it is of course possible to oppose Kim by denying Alexander's Dictum and claiming that epiphenomenal properties can be real. At first, this may not seem like such a bad alternative. However, even if we forget the metaphysical problems associated with epiphenomenalism, we would face some drastic consequences. Let us consider the possibility that qualia are epiphenomenal. This would mean that my itching has no role in causing my scratching, my feeling sick has no role in causing my vomiting, my consciously perceiving a bug has no role in causing me to smash it, and so on. This is highly implausible or even absurd.

7.2. The debate on downward causation[21]

The term "downward causation" comes from Donald Campbell (1974). According to Campbell, in hierarchically organized biological systems, downward causation from higher levels of organization to lower levels takes place. Especially, in natural selection, higher-level laws restrain and "edit" the processes at lower levels. As Campbell himself admits, the expression he has chosen to describe this phenomenon is a little bit misleading: "'Downward causation' is perhaps and awkward term … The 'causation' is downward only if substantial extents of time, covering several reproductive generations, are lumped together as one instant for purposes of analysis. In the 'instantaneous' causation of the older philosophical analyses of physics, no such direction is present." (1974, 180-181). Therefore, Campbell's downward causation has little to do with the sense in which the expression is used in current philosophy.

On the other hand, the distinguished brain scientist Roger Sperry has tirelessly argued in favour of downward causation, much in the sense that interests us here[22]. What Sperry has tried to show is that "emergent" mental or conscious phenomena have a determinative influence and effect on the underlying neurophysical processes. Most of the discussion on Sperry's theory has centred round a seemingly simple example that he has given in "A Modified Concept of Consciousness" (1969):

> The subjective mental phenomena are conceived to influence and to govern the flow of nerve traffic by virtue of their encompassing emergent properties. Individual nerve impulses and other excitatory components of a cerebral activity pattern are simply carried along or shunted this way and that by the prevailing overall dynamics of the whole active process (in principle – just as drops of water are carried along by a local eddy in a stream or the way the molecules and atoms of a wheel are carried along when it rolls downhill, regardless of whether the individual molecules and atoms happen to like it or not). (1969, 532).

Sperry has later emphasized that the wheel rolling downhill is not just an analogy or metaphor: "it is a direct, simple, objective, physical example of macro-

[21] This section is largely based on Klee (1984) and Stephan (1999, 201-210).
[22] However, Sperry prefers the term "macro-determination" to "downward causation".

causation illustrating the universal principle of how the emergent properties of an entity as a whole exert downward causal control over the parts and the trajectories of the parts through space and time without interfering with the causal interactions of the subentities at their own lower levels" (1986, 266).

Sperry's point seems to be that there are numerous cases where the emergent properties of complex systems have a determinative effect on the underlying processes. The control that emergent mental properties exert on neurophysical events is just one example of this downward causation that is to be seen everywhere in nature.

The main defects of Sperry's theory have been pointed out by (among others) Robert Klee (1984). The motion of the wheel as a whole does seem to affect an individual molecule, but this influence is effected through the structural connections this molecule has to neighbouring molecules in the wheel. The situation can be wholly accounted for with micro-connections, we don't need to postulate any macro-to-micro determination. It would seem that the situation is relevantly different with regard to mental properties and neurophysical events, but according to Sperry, both situations are examples of the same universal downward causation.

What we would need is a plausible mechanism of macrodetermination, but Sperry fails to present one. In addition, Sperry's accounts of the nature of macrodetermination sometimes seem to conflict each other. In one article, he writes: "From the start I have stressed consistently that the higher-level phenomena in exerting downward control do *not disrupt* or *inter*vene in the causal relations of the lower-level component activity. […] There need be no 'reconfiguring' of molecules relative to each other within the wheel itself. However, relative to the rest of the world the result is a major 'reconfiguring' of the space-time trajectories of all components in the wheel's infrastructure." (1991, 230). However, in an earlier article, there is the following description: "[C]onscious phenomena as emergent functional properties of brain processing exert an active conrtol role as causal determinants in shaping the flow patterns of cerebral excitation. […] One can compare the rolling wheel to an ongoing brain process or a progressing train of thought in which the overall properties of the brain process, as a coherent organizational entity, determine the timing and spacing of the firing patterns within its neuronal infrastructure." (1980, 201).

Achim Stephan (1999, 201-210) has tried to make sense of Sperry's ideas and has found two different ideas of downward causation implicit in his works. *Weak*

macrodetermination is effected by microreducible systemic properties, does not lead to direct reconfiguration of system components and its mechanism is in principle explainable. *Strong macrodetermination* is effected by irreducible emergent properties, it leads to direct reconfiguration of system components and its mechanism is not even in principle explainable. Sperry's wheel example and all other cases of intra-physical downward causation would seem to fall under the first type, but psychophysical downward causation under the second type.

Weak macrodetermination does not involve any "real" downward causation from systemic properties to system components and it is compatible with reductive physicalism. On the other hand, strong macrodetermination is the form usually associated with emergentism, it is not compatible with reductive physicalism, and it leads to a whole lot of trouble. There are absolutely no examples in nature of the disturbance of the law-like behaviour of system components through causal influence from systemic properties. In addition, if mental properties have causal influence on neurophysical processes, then the causal closure of the physical realm is breached, and physical sciences cannot, even in principle, explain all physical events. Therefore, both empirical evidence and metaphysical reflections are against strong macrodetermination. The only alternative left for an emergentist would seem to be epiphenomenalism, and this option is quite undesirable for the reasons discussed in the previous section. In the next section, I will go through the main argument against mental causation in more detail. In the last section, I will try to see if there is any way out for the poor emergentist.

7.3. Kim's Argument

Kim has presented his argument against mental causation in slightly different ways in the last ten years or so (see for example Kim 1992, 1993, 1998, 2002). The account here is based on the latest formulations (1998, 2002). A similar argument works also against downward causation in general (Kim 1999).

Kim calls his argument "the supervenience argument", as it works against all theories that accept mind-body supervenience. The argument is based on certain principles that together make trouble for mental causation. The first of these is the

principle of the causal closure of the physical domain that can be stated as follows (2002, 276):

> *The Causal Closure of the Physical Domain*: If a physical event has a cause at *t*, then it has a physical cause at *t*.

In terms of explanation: If a physical event has a causal explanation, it has a physical causal explanation. This principle states that physics is causally and explanatorily self-sufficient: There is no need to go outside the physical domain to find a cause or a causal explanation of a physical event. However, the principle does not rule out nonphysical causes or causal explanations of physical events. For this we have another principle (2002, p. 276):

> *Principle of Causal Exclusion*: If an event, *e*, has a sufficient cause, *c*, at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination).

The expression "genuine case of causal overdetermination" refers to cases like two bullets hitting the heart at exactly the same time, both causing death. For the purposes of the argument, it is also convenient to have a generalized and slightly stronger version of this principle (2002, p. 277):

> *Principle of Determinative/Generative Exclusion*: If the existence of an event *e*, or an instantiation of a property *P*, is determined/generated by an event *c* – causally or otherwise – then *e*'s occurrence is not determined/generated by any event wholly distinct from or independent of *c* (unless this is a genuine case of overdetermination).

This broadens the principle to all cases of generation and determination, whether causal or of another kind. According to Kim, the fundamental rationale for the broader principle is the same as for the causal exclusion principle, and anyone who finds the causal principle plausible should also find the broader principle equally plausible.

It is easy to see that these principles generate trouble for anyone who accepts mind-body supervenience and wants to hold on to mental causation. Let us suppose that a mental event, an instantiation of mental property *M*, causes another mental property *M** to be instantiated. This is perfectly consistent with the doctrine of physical causal closure. However, mind-body supervenience says that this instantiation of mental property *M** supervenes on a concurrent instantiation of a physical property *P**. This means that given that *P** is instantiated on this occasion, *M** must necessarily be instantiated on this occasion. That is, the *M** instance is wholly dependent on, and is generated by, the *P** instance.

At this point we are reminded of the exclusion principle. Is the occurrence of the *M** instance due to its supposed cause *M* or its supervenience base *P**? The exclusion principle states that it must be one or the other. Under mind-body supervenience, *M** occurs because its supervenience base *P** occurs, and as long as *P** occurs, *M** must occur no matter what has preceded this *M** instance – in particular, regardless of whether or not an instance of *M* preceded it. *P** alone seems fully responsible for the occurrence of *M**. Given all this, it seems that the only way to reconcile the two causal/generative claims is to say that *M* caused *M** by causing *P** to be instantiated. More generally, it seems that there is a principle involved here that states that *to cause a supervenient property to be instantiated, you must cause one of its base properties to be instantiated.*

At any rate, we know now that the *M* instance must cause a *P** instance. This is an example of mental-physical causation. So what the argument has shown so far is that *mental-to-mental causation presupposes mental-to-physical causation*, and the question is now whether it is possible to make sense of mental-to-physical causation. This is where the principle of causal closure kicks in: the *P** instance must have a physical cause, say *P*. This means that the *P** instance has a physical cause *P* and a mental cause *M*, and the exclusion principle states that one of these must go. Obviously *M* must go, because if *M* were the only cause of *P**, it would violate the physical causal closure. Therefore, *M* cannot be the cause of *M** or of anything else. This holds for all mental properties, and we have the striking conclusion that, under mind-body supervenience, mental properties are causally impotent. Kim has summarized this all as follows (2002, 278):

> *The Problem of Mental Causation*: Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (1) physical causal closure, (2) exclusion, (3) mind-body supervenience, and (4) mental/physical property dualism.

According to Kim, physical causal closure and mind-body supervenience are among the inescapable commitments of all physicalists. The exclusion principle is a general metaphysical constraint that may not seem so indisputable, but even if we abandon it, all we can get is massive overdetermination of physical events by mental events, which is highly implausible.

This leaves only mental/physical property dualism as the one to go. Abandoning it would mean embracing reductionism, which may not seem like a bad alternative, but we have seen in the last chapter that there are good reasons to believe that qualia are fundamentally irreducible. It seems like the problem of mental causation is insolvable for a certain class of properties, that is, phenomenal mental properties or qualia.

This is not just a problem of emergentism. Referring to Schopenhauer, Jaegwon Kim has called it a "world-knot" (Weltknoten), a problem that has eluded our best intellectual efforts and that might be ultimately insoluble (2002, 271). It may arise in every conceptual framework that does justice to the way we subjectively experience the world. In the next section, I will look at some recent reactions to Kim's argument.

7.4. Reactions to Kim's Argument

Kim (1999, 33) himself has proposed that we may save downward (and mental) causation by giving it a conceptual interpretation. On this approach, the same cause can be described in different languages or in terms of different concepts. A single causal relation can be described in different ways, for example in a language referring to mental properties and in a physical language. This would at least save *downward causal explanation*. However, this approach does not solve the ontological part of the problem. With merely conceptual causation, we still have to face epiphenomenal qualia.

Stephan's answer to Kim's problem is to interpret mental causation as *supervenient causation* (1999, 210-218). The idea of supervenient causation is roughly this: *F* superveniently causes *G* iff *F* supervenes on a physical property *m*(*F*) and *G* supervenes on a physical property *m*(*G*) and *m*(*F*) causes *m*(*G*)[23]. According to Stephan, we can thus grant mental properties a causal role without having to deny physical causal closure. However, the problem with this approach is that in supervenient causation, mental properties have causal powers only in virtue of the physical properties on which they supervene – the physical properties do all the work and fully explain the causal relations. It would seem that the mental properties still don't have any *real* causal powers and we still have to face epiphenomenalism. (See also Pihlström 2002, 144-147.).

Barry Loewer (2001) has claimed that Kim is thinking of causation as a relation in which the cause *generates* or *produces* the effect, and this is what makes the problem seem insoluble. The fundamental laws and facts of physics do not mention causation at all. What physicalism is committed to is just that whatever causal relations there are, they must supervene on physical laws and facts. All we can get and all we need is a conception of causation spelled out in terms of difference making and counterfactuals. With this, we can save mental causation – for example, we can say that my itching caused my scratching because if I hadn't felt an itch, I wouldn't have scratched. There are well-known problems with counterfactual accounts of causation, but Loewer tries to overcome them by modifying David Lewis's counterfactual theory of causation.

I will not go to the details of Loewer's account of causation, for I don't think this approach removes the problem. The claim that causal relations are only supervenient on physical laws and facts does not remove the need for metaphysical explanations for these relations. Counterfactual accounts of causation merely state these relations and do not explain them. For example, it is not enough to say that itching is causally related to scratching, this relation needs to be metaphysically explained. (See Kim 1998, 67-72).

El-Hani and Pihlström (2002a, 2002b) have pointed out that most of the participants in the current debate on emergence are firmly committed to metaphysical realism. The central idea of metaphysical realism is that there is a way the world is

[23] If we assume that properties themselves don't have any causal powers, we should say something like "an instantiation of F superveniently causes an instantiation of G iff…" etc.

independent of our minds, language and representations. However, this position has been strongly criticized by philosophers like Davidson, Putnam and Rorty. The emergentists should consider pragmatist alternatives to metaphysical realism, as this could show a way out of many problems, also the problem of mental causation. For example, a pragmatic pluralist view on the mind would allow not only a physical/biological perspective on mentality but also more "humanistic" perspectives.

With this approach, we could also use several different notions of causation. Kim considers mental causation as an instance of normal efficient causation, and this makes the problem seem insoluble. However, with a pluralistic approach on causation that allows different notions of causation for different purposes, the problem can be avoided. For some purposes we need the notion of efficient causation, but for downward (and mental) causation, we need a different notion.

Another solution to Kim's problem proposed by El-Hani and Pihlström (see also Pihlström 2002) is to stick with the natural-scientific understanding of causation as efficient causation but to claim that such a causal vocabulary is inappropriate for adequately understanding human mentality and agency. We should not try to argue that the "higher levels" of human mentality are causally efficacious, but rather accept that human life cannot be fully accounted for in terms of causal concepts. One more proposition of these writers is that perhaps the notion of causation should be relativized to the human mind, in the spirit of Kant.

I cannot go deeper into the details of this pragmatist approach to the problem of emergence. However, I believe that philosophers like El-Hani, Pihlström and Loewer are right in arguing that Kim's ontological presuppositions are far from self-evident and should be critically examined. The problem of mental causation may indeed arise in every conceptual framework that does justice to the way we subjectively experience the world, but Kim's argument is certainly tied to certain metaphysical presuppositions.

8. Conclusion

As I wrote in the introduction, I had two main goals while writing this thesis. The first was to find out what emergence is and what the history of emergentism is. The other was to find out the significance of the concept of emergence for contemporary philosophy of mind. I hope I have at least in some ways reached these goals.

In Part I, I went through the history of emergentism and analyzed the different forms of emergentism, based on Achim Stephan's classification. In this classification, emergentism is divided into *weak*, *diachronic*, *structural* and *synchronic* emergentism. I argued that the only form of emergentism really significant for the philosophy of mind is synchronic emergentism. The defining characteristic of synchronic emergentism is *irreducibility*, in the sense of *behavioural* or *functional* unanalyzability. In the historical overview, main emphasis was on C. D. Broad's *The Mind and its Place in Nature* that is by far the most important work of the old emergentism.

The core of Part II was the argument of the *new qualia emergentism*. The argument was this:

1. Reduction is explanatory only when it reveals the causal mechanisms that fill the causal role of the reduced property, that is, only if it follows the functional model of reduction.
2. The essential characteristics of qualia are not captured by their causal roles.
3. Properties that cannot by explanatorily reduced are emergent properties.

Therefore, qualia are emergent properties.

I believe this is quite a strong argument. Perhaps its weakest point is premise 1, because the functional model is still quite new and just one from several different models of reduction. In the cases of temperature and the gene the model works well, but what about reduction of, say, economics or social sciences? Does the model adequately describe the actual process of reduction in the sciences? Are their other models of reduction that could be called explanatory?

If we accept the argument and embrace emergentism, we have to face the problem of mental causation, which is one of the most puzzling problems in current philosophy. Kim's supervenience argument seems to show that mental causation is impossible if mental properties are distinct from physical properties. However, the argument is based on some strong metaphysical assumptions, and examining them critically could show a way out. On the other hand, it may be that there simply is no solution to this problem, and we should just accept the fact that solving it is beyond the limits of our intellectual capabilities as human beings.

All in all, it is certainly worthwhile to continue research on emergence and reduction. There are many interesting issues that have not yet been adequately dealt with. I have already mentioned the problem of reduction and the functional model of reduction. Kim's argument against mental causation will probably be debated for a long time. The history of emergentism, especially before and outside British Emergentism, is still not very well known. And perhaps most importantly, the metaphysical assumptions that underlie the whole debate have not been adequately discussed. I believe that the real debate on emergence has just begun.

REFERENCES

Alexander, Samuel (1920) *Space, Time, and Deity*. New York: Dover Publications.

Armstrong, David M. (1968) *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

Beckermann, Ansgar (1992a) "Introduction – Reductive and Nonreductive Physicalism" in Beckermann, A., Flohr, H., & Kim., J. (eds.) (1992), 1-21.

Beckermann, Ansgar (1992b) "Supervenience, Emergence and Reduction", in Beckermann, A., Flohr, H. & Kim, J. (eds.) (1992), 94-118.

Beckermann, Ansgar (1997) "Property Physicalism, Reduction and Realization" in Carrier, M., & Machamer, P. K. (eds.) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh: University of Pittsburgh Press, 303-321.

Beckermann, A., Flohr, H., & Kim, J. (eds.) (1992) *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: Walter de Gruyter.

Boogerd, F. C., Bruggeman, F. J., Richardson, R. C., Stephan, A., & Westerhoff, H. V. (forthcoming) "Emergence and its place in nature: A case study of biochemical networks", *Synthese* (forthcoming).

Broad, Charles Dunbar (1921) "Prof. Alexander's Gifford Lectures", *Mind* 30, 25-39, 129-150.

Broad, Charles Dunbar (1925) *The Mind and its Place in Nature*. London: Routledge & Kegan Paul.

Broad, Charles Dunbar (1959) "Autobiography" in Schilpp, Paul Arthur (ed.) (1959). *The Philosophy of C. D. Broad*. New York: Tudor Publishing Company, 3-68.

Bunge, Mario (1977) "Emergence and the Mind". *Neuroscience* 2, 501-509.

Campbell, Donald T. (1974) "'Downward Causation' in Hierarchically Organised Biological Systems" in Ayala, F. J. & Dobzhansky, T. (eds.) *Studies in the Philosophy of Biology*. Berkeley, Los Angeles: University of California Press, 179-186.

Caston, Victor (1997) "Epiphenomenals, Ancient and Modern", *Philosophical Review* 106.

Churchland, Paul (1985) "Reduction, Qualia and the Direct Introspection of Brain States", *The Journal of Philosophy* 82, 8-28.

Churchland, Paul (1988) *Matter and Consciousness*. Revised Edition. Cambridge, MA, London: MIT Press.

Cummins, Robert (1983) *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Davidson, Donald (1970) "Mental Events" in Foster, Lawrence & Swanson, J. W. (eds.), *Experience and Theory*. London: Duckworth. Reprinted in Davidson, Donald (1980) *Essays on Actions and Events*. Oxford: Clarendon Press, 207-227.

Dupré, John (2000) "Reductionism" in Newton-Smith, W. H. (ed.), *A Companion to the Philosophy of Science (Blackwell Companions to Philosophy)*. Oxford, Malden: Blackwell, 402-404.

El-Hani, Charbel N. & Pihlström, Sami (2002a) "Emergence Theories and Pragmatic Realism", *Essays in Philosophy* 3(2). Available online at http://www.humboldt.edu/~essays

El-Hani, Charbel N. & Pihlström, Sami (2002b) "A Pragmatic Realist View of Emergence", *Manuscrito* 25, 105-154.

Fodor, Jerry (1974) "Special Sciences", *Synthese* 28, 97-115.

Gertler, Brie (1999) "A Defense of the Knowledge Argument", *Philosophical Studies* 93, 317-336.

Gillett, Carl & Loewer, Barry (eds.) (2002) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

Hempel, Carl Gustav & Oppenheim, Paul (1948) "Studies in the Logic of Explanation", Philosophy of Science 15, 135-175. Reprinted in Hempel, C. G. (1965) *Aspects of Scientific Explanation.* New York: Free Press, 245-295.

Hooker, C. A. (1981) "Towards a General Theory of Reduction", *Dialogue* 20, 38-60, 201-236, 496-529.

Horgan, Terence (1993) "From Supervenience to Superdupervenience". *Mind* 102.

Jackson, Frank (1982) "Epihenomenal Qualia", *Philosophical Quarterly* 32, 127-136.

Jackson, Frank (1986) "What Mary Didn't Know", *The Journal of Philosophy* 83, 291-295.

Kim, Jaegwon (1992) "'Downward Causation' in Emergentism and Nonreductive Physicalism", in Beckermann, A., Flohr, H. & Kim, J. (eds.) (1992), 119-138.

Kim, Jaegwon (1993) “The Nonreductivist’s Troubles with Mental Causation” in Heil, J., & Mele, A. (eds.) *Mental Causation*. Oxford: Clarendon Press, 189-210.

Kim, Jaegwon (1996) *Philosophy of Mind*. Boulder: Westview Press.

Kim, Jaegwon (1997) “Supervenience, Emergence, and Realization in the Philosophy of Mind” in Carrier, M., & Machamer, P. K. (eds.) *Mindscapes: Philosophy, Science, and the Mind*. Pittsburgh: University of Pittsburgh Press, 271-293.

Kim, Jaegwon (1998) *Mind in a Physical World: An Essay on the Mind-body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kim, Jaegwon (1999) “Making Sense of Emergence”, *Philosophical Studies* 95, 3-36.

Kim, Jaegwon (2002) “Mental Causation and Consciousness: The Two Mind-Body Problems for the Physicalist” in Gillett, C., & Loewer B. (eds.) (2002), 271-283.

Klee, Robert L. (1984) “Micro-Determinism and Concepts of Emergence”, *Philosophy of Science* 51, 44-63.

Levine, Joseph (1983) “Materialism and Qualia: The Explanatory Gap”, *Pacific Philosophical Quarterly* 64, 354-361.

Levine, Joseph (1993) “On Leaving Out What It’s Like” in Davies, M., & Humphreys, G. W. (eds.) (1993) *Consciousness*. Oxford, Cambridge: Blackwell, 121-136.

Lewis, David (1966) “An Argument for the Identity Theory”, *The Journal of Philosophy* 63, 17-25.

Lewis, David (1988) “What Experience Teaches”, *Proceedings of the Russellian Society*, Sydney. Reprinted in Lycan, William G. (ed.) (1990) *Mind and Cognition*. Oxford, Cambridge: Blackwell, 499-519.

Lloyd Morgan, Conwy (1923) *Emergent Evolution*. London: Williams and Norgate.

Loewer, Barry (2001) “Review: Mind in a Physical World, Jaegwon Kim”, *The Journal of Philosophy* 2001, 315-324.

McLaughlin, Brian P. (1992) “The Rise and Fall of British Emergentism” in Beckermann, A., Flohr, H., & Kim, J. (eds.) (1992), 49-93.

Mill, John Stuart (1843) *System of Logic. Ratiocinative and Inductive.* Collected Works, Volumes 7 and 8 (1996). Toronto: University of Toronto Press.

Nagel, Ernest (1961) *The Structure of Science*. London: Routledge & Kegan Paul.

Nagel, Thomas (1974) “What Is It Like to Be a Bat?”, *The Philosophical Review* 83, 435-450.

Nemirow, Laurence (1980) “Review: Mortal Questions, Thomas Nagel”, *The Philosophical Review* 89, 473-477.

Nida-Rümelin, Martine (1995) ”What Mary Couldn’t Know: Belief About Phenomenal States” in Metzinger, Thomas (ed.) (1995). *Conscious Experience*. Paderborn: Schöningh, 219-242.

O’Connor, Timothy & Wong, Hong Yu (2002) “Emergent Properties”, *The Stanford Encyclopedia of Philosophy (Winter 2002 Edition)*, Edward N. Zalta (ed.), URL=<http://plato.stanford.edu/archives/win2002/entries/properties-emergent/>

Oppenheim, Paul & Putnam, Hilary (1958) “Unity of Science as a Working Hypothesis” in Feigl, H., Scriven, M., & Maxwell, G. (eds.) *Minnesota Studies in the Philosophy of Science*, Volume II. Minneapolis: University of Minnesota Press, 3-36.

Pepper, Stephen C. (1926) “Emergence”, *Journal of Philosophy* 23, 241-245.

Pihlström, Sami (2002) “The Re-Emergence of the Emergence Debate”, *Principia* 6, 133-181.

Popper, Karl (1977) *The Self and its Brain*. Part I (of the book written with John Eccles). Heidelberg, Berlin, London, New York: Springer International.

Putnam, Hilary (1967) “Psychological Predicates” in Capitan, W. H. & Merrill, D. D. (eds.) (1967), *Art, Mind and Religion*. Pittsburgh: University of Pittsburgh Press, 37-48. Reprinted as “The Nature of Mental States” in Putnam, Hilary (1975), *Mind, Language and Reality*. Cambridge, MA: Cambridge University Press, 429-440.

Schmitt, Frederick F. (1995) “Naturalism” in Kim, J., & Sosa, E. (eds.) *A Companion to Metaphysics (Blackwell Companions to Philosophy)*. Oxford, Cambridge, MA: Blackwell, 343-345.

Sellars, Roy Wood (1922) *Evolutionary Naturalism*. New York: Russell & Russell.

Sperry, Roger (1969) “A Modified Concept of Consciousness”, *Psychological Review* 76, 532-536.

Sperry, Roger (1980) “Mind-Brain Interaction: Mentalism, Yes; Dualism, No”, *Neuroscience* 5, 195-206.

Sperry, Roger (1986) “Macro- Versus Micro-Determinism”, *Philosophy of Science* 53, 265-270.

Sperry, Roger (1991) “In Defense of Mentalism and Emergent Interaction”, *The Journal of Mind and Behavior* 12, 221-246.

Stephan, Achim (1992) “Emergence – A Systematic View on its Historical Facets”, in Beckermann, A., Flohr, H. and Kim, J. (eds.) (1992), 25-48.

Stephan, Achim (1998) “Varieties of Emergence in Artificial and Natural Systems”, *Zeitschrift für Naturforschung* 53c, 639-656.

Stephan, Achim (1999) *Emergenz: von der Unvorhersagbarkeit zur Selbstorganisation*. Dresden, München: Dresden University Press.

Stephan, Achim (2002) “Emergentism, Irreducibility, and Downward Causation”, *Grazer Philosophische Studien* 65, 77-93.

Stoljar, Daniel (2001) “Physicalism”, *The Stanford Encyclopedia of Philosophy (Spring 2001 Edition)*, Edward N. Zalta (ed.),
URL = <http://plato.stanford.edu/archives/spr2001/entries/physicalism/>.