View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Edinburgh Research Archive

Under the Spell of Multiple Realizability

- A defence of reductionism in mind studies -

Vincenzo Guido Fiore

Master of Science by Research

The University of Edinburgh

academic year 2006-2007

Index

When abstraction sets to killing you,

you've got to get busy with it

The Plague (La Peste),
by Albert Camus, 1947

Introduction

The objective of this dissertation is to question the plausibility of the multiple realizability theory and the method of analysis it entails, opening to a theoretical alternative which can be grounded on a different methodology.

Forty years have passed since Hilary Putnam's article "Psychological Predicates" (1967) in which he formalized his famous argument in favour of the multiple realizations of mental states. His declared intention was to attack the theory of behavioural disposition and the identity theory, presenting a new approach based on the identification between mental processes and the ones performed by a specific virtual machine: the probabilistic automaton. Beside these goals, it is likely that Putnam also intended to challenge all the reductionist approaches to the study of mind denying any other relation between mental and physical states but the contingent one. As a result, the new functionalist approach could focus directly on the information processes implemented by biological systems making of the "sets of instructions" and the "functional states" the new subjects of mind analysis.

A few years later, Fodor (1974) moved farther starting from Putnam's position, making the irreducibility of "special sciences" (i.e. every psychological or social science) the necessary consequence of the assumption of the multiple realizability of mental states. The generalised version of the theory made the mental states multiply realizable by the same correlate structure over time, bringing forth the rejection of all possible fixed identities, even when established between the processes in an organism and a single "best description" of the processes in the correspondent virtual machine. The concept of a "functional state" became more abstract to escape all forms of identity, causing a shift that moved the new generalised version of the theory far from Putnam's early formalization, altering the reference of the main object of the multiple realizations, which became the vague "function".

In the last four decades, the theory has evolved significantly both in the way it has been used as an argument against reductionism and in the reasoning used to sustain it. In this dissertation, I propose to distinguish two main periods: the first lasts for nearly three decades and it is characterised by a general acceptance of the main argument for the multiple realizability, focusing the controversy on its implications on the problem of reduction –Fodor's autonomy of special sciences and Jaegwon Kim's (1989, 1992) local reductionism are two examples of this kind analysed in the text–.

On the contrary, in the nineties, and more frequently in the last ten years, a new phase has started, which is centred on two foci: on one hand there is a strong challenge to the conceivability of the main argument itself (Zangwill 1992, Shapiro 2000). On the other hand some philosophers are now stressing the failures in the predictions implied by the generalised theory: neuroscience is having a great impact in the study of psychological and social sciences and most of its studies are successfully carried out assuming an across-species neural substrate that may be reconciled with the multiple realizability theory or may deny it, depending on the grain of analysis adopted (Bechtel and Mundale, 1999).

Nonetheless, neuroscience findings do not lead by themselves to abandon the hypothesis that mental states are multiply realised in several neural correlates. It is sufficient to consider that, using a fine grain of analysis, even if we consider neural systems belonging to the same species, it is not possible to find two perfectly identical ones: it is not then a surprise that neuroscience findings have been used both to sustain and to oppose the hypothesis for reduction, when they are asked to give an account of multiple realizations.

In this dissertation I propose my contribution to the controversy assuming a connectionist perspective: an approach that has been so far disregarded in this particular context. My strategy consists in accepting the idea that it is useful to analyse organisms as information processing systems, but I will claim that Putnam's hypothesis relies on a double misconception. Appealing to Copeland's (2002) article about the Church-Turing thesis, I will try to demonstrate in the second chapter that these false beliefs concern which type of mathematical functions can be considered computable by which kind of device. Namely, philosophers usually forget that Turing machines cannot replicate every possible mathematical function and it is yet to be demonstrated if they are able to perform the specific mathematical functions computed by neural systems.

Such a result may be only achieved with an empirical study of parallel system processes and I am convinced that, from a mathematical perspective, these two classes of systems (the first serial, the second parallel) share little or nothing, so that one cannot process the information as the other does it. If the organisms do not process information in the way probabilistic automata do it, then the identification of the mental state with a functional state collapses, making the hypothesis of multiple realizations of the mental unlikely to be verified.

The early version can be questioned with a mathematical approach, but the generalised version escapes this "fate" thanks to the abstract (or elusive) reference attributed to the term of its investigation, the "function". I will then rise a series of objections against the possibility of identifying the function realised by a system with the behaviour generated by a mental state (in terms of evolutionary psychology) or with the "main task" accomplished by a generic device. I see this defining strategy as affected by a general vagueness due to its aim for abstraction and to its dependence from the point of view of the describer of the function. Nevertheless, even if it were possible to establish a shared method for the identification of the functions in the generalised version of the theory, this strategy would lead inevitably to a partial description of the phenomenon analysed.

In the early version of the theory, Putnam could reach a complete description of the processes realized by an organism thanks to the identification of these with the ones performed by the correspondent virtual machine. The generalised version of the theory cannot reach the same result and its analysis focuses on single parts of the whole system. Generally speaking, my thesis is that partial descriptions may be useful in some contexts (i.e. if we do not have direct access to the phenomenon analysed and its complete description), but they fail to give us exhaustive explanations and should be considered weak as the ground for an explanatory model.

Moreover, partial descriptions are likely to affect negatively our understanding of complex phenomena: it is usually taken for granted that a part of the system, once separated from the whole, would still be analysable as performing the same processes and functions. A similar assumption is often mistaken as plausible concerning the data analysed: each selection may lead to a different understanding of the processes that generate those particular data.

The first assumption is easily falsified in this context on the basis of the mathematical description of the processes performed by parallel systems: every attempt to isolate a part of the system leads to the generation of a different, new system. The *double-slit experiment*, which is famously used in physics to show the double nature of matter, is then used to exemplify how the selection of the data gives the observer two representations of the phenomenon leading to incompatible models.

In conclusion, the thesis here supported is that the tool of neural computation may lead us in the near future to complete descriptions relying on the "mathematical functions" implemented by a system (i.e. mathematical description of the processes

performed by the system) and these can only be formalized studying the physical matter that generates them.

As far as multiple realizability is concerned, I consider it a necessary tool rather than an explanatory theory: whenever the processes in a system “A” are not accessible, they can be investigated by means of the assumption that another system “B”, whose processes are accessible, is partially realizing the same (or similar) processes. The reasoning will lead to an analogy, but the descriptions achieved should not be confused with those which refer directly to “A”, because it is necessary to bear in mind that they refer to a system that is *assumed* to be similar. Consequently, the assumption itself (and all the descriptions it allows) can be falsified the very moment the system “A” becomes accessible thanks to new tools of analysis which are bringing forth new explanatory models.

Chapter 1

1. Putnam's overwhelmingly plausible idea. In the late sixties Hilary Putnam published a series of papers introducing the thesis of multiple realizability in the theory of philosophy of mind. Above all, "Psychological Predicates" (1967)[1] is commonly recognised as the pivotal attempt to formalize this theory giving a new explanation of the relation between the mind and the body (or, as Putnam himself stated years later, resuming and modifying the ancient Aristotelian hylomorphism).

At the time Putnam wrote his articles, Feigl and Smart's psychoneural identity theory was widely accepted: their "type physicalism" supported a type-to-type identity between brain states (or sometimes the theoretical assumption of a physical "C-Nerve fibre activation") and mental states which was only challenged by anti-physicalist position. Putnam's arguments radically changed this dichotomy, focusing for the first time on the processes that give origin to the mental state and famously associating the mind with a virtual machine. Putnam's main innovation, particularly if his theory is considered in relation to traditional approaches to the study of the mind –namely, physicalist monism and Cartesian dualism–, consists in allowing a physicalist position that unbinds the study of the mind from the study of the matter that generates it[2]. It is due to this innovation if the theory could resist for a long period to the challenges coming from the field of neuroscience, maintaining its philosophical appeal in spite of the numerous changes and findings that have characterised the brain studies in the last four decades.

Putnam starts his article with an analysis of the usual criticisms against type-physicalism concluding that they fail to reach their target because they are wrongly addressed. In order to show the reasons of these failures, according to the author it is first necessary to underline the implications an identity has when it is used within analytical philosophy. The statement "*A* is *B*" (which stands for "being *A* is being *B*")

[1] The article was first published in *Art, Mind and Religion*, eds. W. H. Capitan and D. D. Merrill (Pittsburgh, Penn.: University of Pittsburgh Press, 1967), pp. 37-48. It was reprinted in 1975 with the new title "The Nature of Mental States" in *Mind, Language and Reality*, pp. 429-440 and the subsequent reprints use the latter title. In this dissertation I will always refer to the 1999 reprinted version for the anthology "Mind and Cognition" edited by W.G. Lycan. See references for details.

[2] In section 3 it will be analysed how this innovation has been questioned by a series of article by Jaegwon Kim. Despite the fact Kim is not alone in his criticism, he is certainly in a minority.

should be considered correct if and only if it follows from what *A* and *B* mean and it must be reductive in order to be informative.

Grounding his analysis on this definition, Putnam rejects the hypothesis that there is a linguistic violation in statements such as "the mental state *A* is a brain state *B*". The hypothesis suggests that the condition of being aware of experiencing a mental state (for instance, pain) does not entail being aware of any correspondent brain state. Putnam does not question this hypothesis, but he claims that the only consequence that can be drawn from it is that the two kinds of knowledge rely on two distinguished concepts. As a matter of fact, granted the previously mentioned definition, every informative identity has to be established between two concepts: supporting this argument against type-physicalism would consequently make all scientific identities unacceptable (e.g. "water is H_2O", "heat is molecular kinetic energy" and so forth).

This is not the only hypothesis the author analyses that should be extended (if properly used) to all scientific statements concerning identities. A second strategy consists in assuming that an identity entails that the properties that characterise the two elements involved, must be synonyms. According to the author, such a use of an identity would lead to make the two notions of "property" and "concept" collapse in a single one. The problem becomes clear once this reasoning is applied to recognised identities in physics: Putnam uses the example of the temperature associated to kinetic energy showing how the statement "A is B" is characterised by a different truth value than "the concept of A is the same concept of B"; namely, the first statement is true, whilst the second is false.

Finally, a third strategy that entails an extension of its conclusion to all scientific identities consists in assuming that the properties involved in an identity invariantly correlate with each other (i.e. an attempt to avoid the controversy of causal relation as well as identities). The author shows how this assumption does not lead to a solution of the problem. A correlation between two properties would simply cause a change in the question from "is *A B*?" to "what makes *A* correlate with *B*?". Furthermore, once the reasoning is extended to every scientific identity, it entails doubting the existence of any causal relation (which is substituted by invariant correlation) or considering that every known relation in physics has been established between causally irrelevant elements.

A third commonly used theoretical hypothesis against the identity theory is based on the assumption that it should be possible to establish a rigid spatiotemporal association between the two properties involved in the identity, in order to have a proper identification. Thus, since the sensation of pain in my arm clearly does not share its location in the space with the status of activation in the brain, the identity "pain in my arm is such-and-such brain state" should not be accepted. This assumption is rejected by Putnam who claims that apparently nobody would be confident of using the same strategy to demonstrate that the surface of the mirror is not actually reflecting lights, despite the fact that we perceive the reflected objects as if they were "behind" the mirror itself.

These analyses of the usual criticism against the identity theory lead Putnam to conclude that one cannot argue with the hypothesis that a mental state and a certain physical state in the brain (or the nervous system) exemplify two different properties. Consequently, since the aim is to discuss what pain "is", the question whether there is a proper identification between the two properties is still open.

At this point, Putnam presents his famous argument:

> Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that any organism (not just a mammalian) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. (Putnam 1967, 31).

Putnam considers obvious that every animal, independently from the species or class it belongs to (i.e. mollusca as well as mammals or reptiles), is capable of being in the particular mental state that corresponds to pain. Furthermore, considering this criterion, the mental state of pain must be identical for every creature (i.e. there is no such a thing as a specific human or mollusc pain) and there is no reason to doubt that even hypothetical creatures like aliens or artificial silicon-based system can be also conceived as characterised by the same mental state of pain.

As a consequence, if pain is a certain C-Fiber activation, then all the creatures in the universe capable of feeling pain should share the same neural structure and the same neural activation at the right moment. Of course even if we consider, as Putnam himself does, that parallel evolution might lead to the same physical structure, once

the argument is extended to every psychological predicate, it becomes overwhelmingly plausible that these multiple realizations of the mental states across species simply cannot be explained by any theory grounded on the identity between mental and physical states. After all, even if parallel evolution could be proved in all known creatures, this would be a weak defence of the theory: the conceivability of artificial intelligent systems not based on biological neural networks but still capable of feeling pain, would definitively discard the identity theory.

It is interesting to stress how Putnam's arguments rely on his use of deductions concerning neurophysiology and neuroanatomy, so that the tools used by reductionist philosophers are overturned against physical reductionism itself. Rather than trying to disprove the identity theory using a priori assumptions, this vivid argument (sometimes referred to as the likelihood argument) makes the theory sound highly implausible.

Putnam's explicit strategy is then to argue that there is another approach to the mind that allows a more plausible explanatory model of what mental states are. To this end, he introduces the concept of the probabilistic automaton, a device similar to a Turing Machine (TM) characterised by a different organization of the transition between states. A TM is a computational -serial- device that is instructed by a program (set of instructions) to process a symbolic input in order to give a symbolic output as a result. A device that simulates any possible TM is called Universal TM (UTM): in this case, there are sets of inputs that are processed by the machine to generate the correspondent sets of instructions. In other words, the UTM is directly programmed by the input which instructs the machine on the processes to apply from that moment onwards.

Given the input, a TM follows its program step by step -and this is the reason why I mentioned before that it is a serial system- in what is called a transition between states. If and only if the correct conditions are verified, the transition necessarily occurs allowing the machine to change its state of origin into a new one, described by its set of instructions: in other words, a TM is characterised by an assignment of probabilities 1 or 0 to every transition. On the contrary, if the instructions allow the machine to change its status from the original one to a series of target ones (with probabilities assigned to each of them) then the machine is called "probabilistic automaton": according to Putnam, this device describes the functional processes of an organism better than a TM.

Putnam considers on one hand the signals generated by the sense organs as a physical realization of the symbolic input of the automaton and on the other hand the observable behaviour performed by the organism as the correspondent output. The probabilistic state transition allows the functional transformation of the input into the output and remains implicit representing the psychological states of the organism. Thus, it is possible to identify a mental state of an organism on the basis of its behaviour, even if the behaviour itself (as every output) is not necessary for the existence of the mental state. The behaviour is rather the manifestation of the inner state of the organism; using a metaphor, the behaviour plays the role of a warning light: it indicates the presence of the correspondent mental state if and only if it is working correctly.

Putnam intentionally specifies his position in relation to the behaviour because he bears in mind the problems generated by the behavioural disposition theory. In contrast with this theory, Putnam's use of observable behaviour does not fail in giving an account of those circumstances that characterise an organism which cannot perform a specific behaviour or any behaviour at all (e.g. due to paralysis). Behavioural disposition is considered by the author as a 'hopelessly vague' and 'clearly false' (Putnam 1967, p. 33) theory: if the behaviour is only referred to what can be observed, then this theory is committed to claim that an animal whose motor nerves have been cut cannot be in the mental state of pain, just because it cannot perform the proper correspondent action. On the contrary, Putnam's focus on the functional organization allows to explain the latter set of circumstances simply with an interruption in the state transitions: the animal will still feel pain, but this mental state cannot be processed until reaching the end of the usual chain of states, which results in the execution of the correspondent action.

Once the focus is on the functional organization implemented by the probabilistic device, it follows that there are infinite systems able to implement these state transitions or singular functional states: the virtual machine is completely independent from any specific physical structure. If the state transitions implemented by two devices are perfectly mirrored, these two systems are *functionally isomorphic*, but this is not a necessary condition. Even if the sequence of states established by the program is different, two systems may still reach the same state in the virtual machine (i.e. the same psychological state) by means of different set of instructions: each state

is independent of the others and it can be reached by the machine by means of different processes.

The instructions the automaton realizes are neither species specific nor they have preferences for any sort of structures, such as, for instance, biological neural networks or microchips. Combining this hypothesis –contingency of the structure that implements the programs– with the features of the UTMs –the instructions of the program themselves may be given as an input–, Putnam concluded few years later with the famous statement: 'we could be made of Swiss cheese and it wouldn't matter' (Putnam 1975, p.291).

In conclusion, the argument may be summarised considering the following relations:

$$\begin{array}{l} A \text{ r } M \\ B \text{ r } M \end{array} \qquad A \neq B$$

A and *B* stand for two neural structures belonging to different species, *M* is used to represent the psychological state and *r* stands for a relation that has to be established. Since *A* and *B* are given as not equal, if we suppose that the relation between them and *M* is an identity, this assumption would lead to the absurd:

$$\begin{array}{l} A = M \\ B = M \end{array} \qquad A = B \text{ and } A \neq B$$

Putnam's proposal is then to assume that *A* and *B* are not the mental state *M* but they *realize* it: the established relation is contingent and there can be infinite systems realizing the same relation with *M* thanks to different sets of instructions for state transitions. As a consequence, the focus changes from the reductionist study of the neural correlate to the functionalist study of the realized functions.

2. Towards the autonomy of "Special Sciences". In the few years following Putnam's formalization, many attempts were made in order to make the identity theory compatible with the commonly recognised plausibility of the argument of multiple realizations of the mental states. A first attempt appealed to the

opportunity to have different kinds of identities depending on specific characteristics of the structure, though preserving the main principle. A frequently quoted example of how physical explanations allow multiple realizations in different microphysical contexts is represented by the identity between molecular kinetic energy and heat. It is known that this identity is true when it is referred to gas, whilst the temperature in solids or in liquids cannot be described exactly in the same way, even if the concept is still valid (e.g. in solids heat is defined as the *maximal* molecular kinetic energy).

Nevertheless, in the case of psychological states, a physical identity still requires the identification of a type-structure or type-property feature of matter that could be identified with all the mental states of the same psychological kind. Putnam's analysis of the chances for parallel evolution perfectly fits with this modified version of the identities, but it has been already stated that such a change in the identity theory would still lack the necessary plasticity that could make it sound plausible. Anyhow, even if we ignore these considerations, it would still be possible to conceive two functionally isomorphic systems grounded on parallel neural systems and serial microchips. Until recent years, this thought experiment has, *per se*, been considered as implying the rejection of the hypothesis.

A more complex path to reconcile type identity theory with the assumption of multiple realizations of mental states may be established thanks to the so-called "functional state identity theory" (from now on, FSIT). This new identity theory can be considered as the direct consequence drawn from Putnam's definition of the probabilistic automaton: Block and Fodor give an accurate analysis of this position in their 1972 article "What Psychological States are not".

According to the authors' description, FSIT theorists consider the input of the device as constituted by sensory stimulation whilst the output represents the instructions for motor transducers, but contrarily to Putnam's proposal, they assume the existence of a unique best description of the transition probabilities between mental states.

The theory *per se* is neutral about what psychological predicates are, in that it simply requires to be applied to a device characterised by the ability to realize them. This new class of identities doesn't imply that a specific device can be considered the only one capable of these realizations: from this point of view, it doesn't matter whether the device in question is a silicon robot, a Martian, a hungry octopus or a

human being; even in the same organism it might be applied to the entire person, a specific neural matter or the soul.

Far from being grounded on ontological choices, Block and Fodor claim that this theory requires two assumptions: first, independently of what mental state are, they are supposed to be in strict correspondence with the virtual machine states. Secondly, there is only one best description of the transition between the states for each device. Nevertheless, considering these premises, the FSIT substantially modifies Putnam's conceivability of the event of multiple realization of a mental state which would then happen if and only if the programs implemented by the devices are functionally isomorphic. Moreover, the theory allows a form of token-identities that, in principle, may become a backdoor for a physicalist attempt to conceive a new version of identities between physical and mental states, maintaining the one-to-one relation established in the FSIT.

According to Block and Fodor, the theory cannot be accepted due to a series of weak points: first, it does not give an account of the distinction between dispositional states (e.g. beliefs and desires or other lasting inclinations) and occurrent states (e.g. sensations and feelings limited to the present time). The FSIT might give an explanation of the occurrent states thanks to the identity one-to-one established with the states of the virtual machine, but it is not possible to describe in this way a disposition because of its lasting feature, unless we don't want to assume that they are composed of occurrent states (a position considered possible by the authors, but dosmissed because it is not empirically grounded and it would require an unjustified conceptual change).

A second criticism, which will come in useful in the last sections, is centred on the fact that it is not possible to give a good explanation of how an organism is capable of simultaneous mental states (a condition which is assumed by the author to be experienced continuously). In order to justify this phenomenon the organism should be described thanks to several probabilistic automata operating in parallel; however, accepting this case would entail a denial of the assumption of the single best (and unique) description, which is instead necessary for the FSIT.

Third, the identity between psychological and machine table states implies that multiple realizability of mental states is only limited to those cases that are consistent with functional isomorphism. A machine table state is identical to another if and only if the inputs that generated them, the output that they are going to create and finally

their probabilistic relations with the other table states are also identical. This assumption generates a strong restriction for multiple realizations and, according to the authors, it ignores the nomological possibility for two functionally identical mental states to have different qualitative contents (expressing different qualia).

Finally, a fourth argument raised by the authors appeals to the insufficient abstraction of the theory in its relation with multiple realizability. This criticism has already been used by the authors against behaviourist and physicalist identities (in a way which is close to Putnam's standard arguments): the problem relies once again on the restriction for multiple realizations of mental states to functionally isomorphic virtual machine. There is no obvious reason to establish a law-like relation between a single mental state and the whole program that becomes necessary in order to realize it, once the FSIT is accepted.

The authors' conclusion is that every type of identity implies a systematic failure: firstly because it cannot give an account of temporal constancy of dispositional psychological states and secondly due to its rigidity in relation to the variety of realizations of the mental states. The sciences grounded on the study of the mind are characterised by a general complexity that avoids the strict one-to-one law imposed by identity theories, independently of the way they are grounded: behaviours (psychological states may have different manifestations), physical states (the problem of the across-species neural correlate) and, it is now necessary to add, virtual machine states (as hypothesised by the FSIT and rejected in the article just analysed).

The assumption that any type of identity is to fail when it is used in relation to the study of psychological predicates commits to claim a special autonomy for those sciences which are connected to the study of the mind: these are indeed the so-called "special sciences". Just two years after the coming out of Block and Fodor's attack on type identities, Jerry Fodor extended Putnam's antireductionist position in "Special Sciences (or: the disunity of science as a working hypothesis)" (1974) bringing forth the "generalised" version of multiple realizability theory.

For the purpose of this dissertation, it is interesting to analyse the way Fodor introduces his argument, before dealing with the main focus of the generalized theory and its consequences for reductionist approaches. First, Fodor draws the relation established between token physicalism, type physicalism and reductionism which are considered three autonomous approaches. According to Fodor:

1) The hypothesis developed in token physicalism is that special sciences describe events which are connected with physical ones by means of "bridge laws". Since the relation supposed to be established between the two kinds of events is an identity, all sciences must be only interested in physical events. This approach should not be confused with the materialist one, which also claims that every possible event is a physical one. Token physicalism allows the existence of events which are not of the physical kind, namely those which are related to the "special sciences": Fodor clearly states this key point in his reasoning drawing the conclusion that this form of physicalism is necessary but it is not sufficient both for type-physicalism and for reductionism.

2) Type physicalism focuses on physical properties rather than on physical events: 'two events will be identical when they consist of the instantiation of the same property by the same individual at the same time' (Fodor, 1974 p. 100). Type physicalism necessarily entails the token one, whilst the vice versa is not implied: the reason is that even a nomologically necessary identity may be contingent and, as a consequence, it would not imply the identity of the properties. It is still possible to conceive bridge laws that bind multiple physical events to a single non-physical one, generating a multiple realization that would deny any possible identity of properties.

3) Reductionism is grounded on the assumption that all natural predicates in any ideally completed science (including the special ones, which at this point loose any feature that makes them "special") may be reduced, by means of perfectly describable laws of reduction, to the natural predicates of an ideally[3] completed physics. According to the author this, the strongest of the three assumptions, implies the unity of sciences[4]. Fodor clearly assumes reductionism entailing token

[3] The reference to the "ideally completed" science underlines Fodor's interest in a criticism of the theoretical possibility of the existence of reductionism, rather than of the practical existence of complete reduction laws in contemporary science. I consider the point of particular interest especially when compared to the author's use of findings coming from neuroscience field (see later on in this section) so that a model on theoretical possibilities of science is grounded on models generated by a science in its first years of development.

[4] For instance, the longest paths are usually conceived to be applied to political science and economics: these are reducible to physics thanks to a series of steps which start from the reduction of the firsts into sociology and goes on through psychology, biology, chemistry and finally gets to physics. The concept of reducibility entails an asymmetric relation between the reduced science and the ones it is reduced to: in agreement with reductionism, all sciences converge to physics, which is the only one characterised by general laws, the *basic* science.

physicalism whilst, as with type physicalism, the vice versa is not a necessary implication due to, once again, the conceivability of contingent identity between physical and special science events.

Starting from Putnam's argument in support of multiple realizations of mental states, until Fodor's analysis of types of physicalism, the theme of contingency recurs as a constant presence against any attempt to establish strict identities. Fodor's objective is not to demonstrate that a single event may not be reduced to some physical event by means of some sort of bridge law. His intention is to demonstrate that it is not possible, starting from this form of reduction, to draw the conclusion that a specific token event related to special sciences is *always* bridged to the same token event in physics. That is to say, all mental states realized in a particular human in a given time are implemented by his or her neural system (assuming a reduction of psychology to neurology), but these are 'nomologically necessary contingent event identities' (Fodor 1974, 101), which do not imply reduction of properties and do not entail the existence of fixed ever-lasting laws *per se*.

If we accept the definition the author uses, we are committed to accept that token physicalism allows the hypothesis that the psychological natural kind is not necessarily co-extensive with the neurological natural kind. If such co-extension could be established it seems likely to hypothesise that it is not law-like: if Putnam is right when he hypothesises the simulation of mental states by probabilistic automata, which are not based on neurological predicates, then, once again, the token identity is characterised by a contingent relation. As a consequence, the other forms of physicalism are 'foredoomed' (Fodor's word) to fail in their attempt to give an account of special science phenomena because they require a relation grounded on necessary identities.

Nevertheless, Fodor does not think the identification between mind processes and virtual machine ones is sufficient to support the autonomy of the special sciences towards reductionism. This is the reason why his *pièce de résistance* consists in appealing to physiological psychologists Karl Lashley: Fodor claims that only coarse correspondences may be established between types of psychological states and types of neurological states[5]: the nervous system of higher organisms is typically able to

[5] A fine correspondence may be established between token-events, but it has been already stated that this assumption would not entail an identity between properties, which is necessary in the path of sustaining reductionism.

accomplish a single psychological task in a wide variety of ways by means of several neurological parts of the whole structure. Herein lies the pivotal difference between Putnam's appeal to contingency and Fodor's one: the role that time plays in the identity relations between psychological and physical states allows to generalise the original theory of multiple realizations on the basis of the neuroscience puzzle about the plasticity of the cortex:

> Imagine a world in which such [type-to-type] correlations are *not* forthcoming. What is found, instead, is that for every n-tuple of type identical psychological events, there is a spatiotemporally correlated n-tuple of type *distinct* neurological events. (Fodor 1974, 106 italics in the original text).

Fodor's argument is intuitively effective: considering the same token-system, at different times it is possible to imagine that psychological states of the same kind may be realised by more than a single type of neural correlate. If the assumption shows to be true, type-identity theory is consequently rejected because it necessarily requires constancy in the elements that generate the identification: time cannot be a variable of the model.

This is not a theoretical speculation: its empirical support consists of the known phenomena of neural degeneracy and plasticity which were both used with good results by Lashley to belittle the numerous attempts to generate both a specific map of the human cortex and a coarse map of cortices that might be meaningful across species. The degeneracy is a characteristic feature of the cortex, particularly studied in human beings, which results in the ability to offset damages in the structure caused by traumas or diseases: different regions have been proved to modify their supposed cognitive function in order to compensate for the missing ones performed by the damaged area. The plasticity refers to the dynamic nature of the whole brain and it has been described in the cortex of superior mammals: even in those subjects which have not been affected by major traumas or diseases, complex biological neural systems change their micro-architecture all along the life of the organism. These micro alterations are caused by genetic dispositions, external stimulations, but they also depend on completely stochastic reasons.

Let us again consider the table I used to summarise Putnam's argument:

$$\begin{array}{l} A \text{ r } M \\ B \text{ r } M \end{array} \qquad A \neq B$$

First, Fodor's generalized argument shows a single token structure considered in two different times in place of the former distinction between the two neural correlates *A* and *B* (i.e. $A = B$). Secondly, M_1 and M_2 represent two different mental states (i.e. $M_1 \neq M_2$) which take the place of the single mental state *M* that characterises Putnam's theory. As a consequence, Fodor's definition of token physicalism would lead to the following contradiction:

$$\begin{array}{l} A = M_1 \ (\text{in } t_1) \\ A = M_2 \ (\text{in } t_2) \end{array} \qquad M_1 = M_2 \text{ and } M_1 \neq M_2$$

The only way to use an identity in place of the generic relation *r* is to assume that physical laws are not constant in time, an assumption which is of course unacceptable. The relations are confirmed to be contingent even when they are applied to a single token device. Therefore, Fodor's generalized multiple realizability supports the idea that it is not possible to reduce psychological states to physical laws: the 'special sciences' reach a higher level of abstraction and must claim their independence.

In the second chapter of this dissertation I will deal with the problems that, in my opinion, are arising from this appeal to a higher level of abstraction: for the time being it is important to point out that in Putnam's early version of the multiple realizability theory there is no mention of the term "function". It is rather the "functional state" of a virtual machine that is the subject of multiple realizations: it simulates in the virtual devices the psychological predicates or mental states which are implemented by biological systems. Even if Putnam does not define clearly what he means by a psychological predicate, it seems plausible that these are synonymous with the functional states, thanks to the hypothesised possibility to simulate the former by means of the latter.

In Fodor's generalization the "psychological functions", or simply "functions", take the place of the functional states with the intention of making the whole theory abstract. In opposition to Putnam's early formalization of the theory, the new

generalised form does not establish an identity between these particular functions and the functional states granting two functionally identical states to have different qualitative contents (as it as been explained dealing with criticisms against the FSIT). What is more, the functional state is defined as a single part in a complex transition process identified, in Putnam's conception, in the probabilistic automata: these virtual machines consist in a complete description of how the input information is computed by means of all the possible transitions between states; consequently, observable behaviour becomes irrelevant for the definition of the state in the machine. Despite the fact they both appeal to the same concept of the virtual machine simulation of the mental, Putnam's descriptive power can hardly be extended to the functions that characterise Fodor's generalised theory. Their definition is usually taken for granted, probably because it is considered of immediate understanding, but, as a matter of fact, Fodor's quest for abstraction commits him to abandon Putnam's clear use of the subject of multiple realizations and paradoxically forces him at least to a partial appeal to the observable behaviour in the attempt to describe the nature of the functions he uses.

Before moving forward to the next sections, it is useful to stress briefly the differences between Fodor's generalised version of the multiple realizability theory and the standard "conceivability argument" (or "argument for imagination", as it will addressed in section four), which is also used in support of the theory and was formalized in the same period.

A classic example of the conceivability theory is represented by Kripke's appeal to God's intervention (Kripke 1972 p.153-154): this is an expedient used in order to let us conceive how the same C-Fibre stimulation may be "used" by God to make the organism feel pain, rather than warmth, or ticklishness or nothing at all. Based on the modal use of the terms *necessity* and *possibility*, Kripke claims that conceivability entails possibility and, as a consequence, that there is no necessary relation (such as a cause-effect) between any C-Fibre stimulation and a mental state.

The conclusion reached by this new path matches the one described as the generalised theory: a single neural correlate may give rise to different psychological states in different times. Nonetheless, these two arguments must not be confused: Fodor's appeal to Karl Lashley's research in the field of neurophysiology allows him to take for granted as scientific evidence the multiple realizations of psychological states in the same neural correlates over time. The use of the words 'Imagine a world

[...]' (Fodor 1974, 106) may be misleading, but it is to notice that Fodor never needs to explain the controversial assumption that conceivability always entails possibility: the mental experiment is used as an example, rather than the ground on which the whole thesis is built. In other words, it helps us to understand the argument, it is not the argument itself.

In the coming sections, I will mainly address my arguments against Fodor's theory and in support of reductionism so that, despite their apparent similarities, it is fair to point out that the objection raised against the generalised theory for multiple realizability cannot be applied to the conceivability argument, and vice versa.

3. The "third way": reconciling reductionism with multiple realizability. For more than two decades after Putnam's formalization of the theory, the field of philosophy of mind was ruled by the arguments for multiple realizations. As a result, reductionist approaches to the mind fell from favour, surrounded by an atmosphere of naivety that made them not even worth being taken into serious consideration.

At the beginning of the article 'The myth of nonreductive materialism', Jaegwon Kim (1989) briefly analyses the way reductionism was perceived by philosophers after the coming of multiple realizability theory: lacking the required plasticity and abstractness, this hypothesis was conceived as monolithic, rigidly bound to the attempt to establish a negative order and, even worst, reductionism started to be considered dogmatic at least every time it was applied to mind studies.

According to the author, the functionalist position coming from the generalised version of the multiple realizability theory, carries on a new physicalism which asserts its freedom from the restrains of reductionism on the basis of a dichotomy. On one side, there is the "ontological physicalism", which claims all existing objects in the space-time are physical. On the other side, the "property dualism" which claims the autonomy of psychological attributes in relation to the physical ones.

Kim's article represents the first attempt to question this specific position, otherwise widely accepted. The author focuses on two topics, but for the purpose of this dissertation I will not deal with the arguments centred on Davidson's "anomalous monism" (which is associated with an eliminativist position), giving here an account of Kim's position in relation to the multiple realizability theory. It is necessary to

point out that the author is not concerned with arguing against the theory itself, which he considers plausible too: he is rather interested in discarding the hypothesis that this theory entails the existence of a physicalist approach which is neither reductionist, nor eliminativist. This intent is clearly stated in the article:

> I will claim that a physicalist has only two genuine options, eliminativism and reductionism. That is if you have already made your commitment to a version of physicalism worthy of the name, you must accept the reducibility of the psychological to the physical, or, failing that, you must consider the psychological as falling outside your physicalistically respectable ontology. (J. Kim 1989, p. 32)

Kim starts analysing Putnam's formalization of the theory, claiming that the main feature of multiple realizability consists in rejecting the hypothesis of the coextension of any physical state with the mental property it realizes. Therefore, the physical state is defined by the author as being "nomologically sufficient" for the realization of the mental, generating the already known relation: "*A* realizes *M*" or, using Kim's symbols, Pi → M.

At this point, Kim notices that both the choices of the examples in the theory and the words used by Putnam in his papers allow thinking that there is a species specificity in the latest relation, so that a better formulation would be:

Si → (M ↔ Pi)

which means that, within a specific structure Si, the relation established between M and Pi (the mental and the physical predicate) is both necessary and sufficient. Kim defines this relation as a 'species-specific biconditional law' (Kim 1989, p.38).

Despite its name, which was explicitly chosen for its simplicity, this law is preferentially applied by the author to structures, rather than to biological species: the difference is slight if we consider that the neural structures of organisms belonging to the same species share a high percentage of their features. Nonetheless, the author states clearly that binding this law to a species would make it fail, due to the neural plasticity and the individual specific variations already highlighted by Lashley[6]. If, on the contrary, the law is applied to structure-specific types, it can be used to give an

[6] See section 2.

account of multiple realizability, and at the same time it entails a one-to-one relation between specific physical structures and the mental states it realizes.

A defence of reductionism in mind studies does not need the author to claim that multiple realizability necessarily entails these *local laws*: it is sufficient that the theory can be considered consistent with them. This consistency would counter the line of argument that leads to the dismissal of every reductionist perspective by allowing these perspectives to be compatible with multiple realizations of mental states.

In conclusion, Kim's local reductionism is the strategy the author proposes to restore cause-effect relations between physical and mental kinds, establishing empirically verifiable laws. Each of these, restrained within specific species or structure, locates a single neural substrate that necessarily accomplishes the analysed cognitive capacity or function.

Three years later, Kim dealt again with his theoretical proposal in the article: 'Multiple realization and the metaphysics of reduction' (1992), which is this time entirely addressed to the argument of local reductionism. In order to clarify his argument, the author uses the example of the stone called jade, establishing an analogy with the commonly used example of pain in psychology. In the past this stone was believed to be a mineral characterised by specific features, the most recognisable being its green colour. Once it has been studied from a microphysical perspective, the stone proved to be a misconceived class composed of two different minerals: jadeite and nephrite which are characterised by similar macro-features and for this reason they were mistakenly considered the same stone.

The question Kim rises is important: is the sentence "Jade is green" still a law? Apparently, it seems that the answer to this question is positive: the law-like expression supports counterfactuals and, using Kim's words, it may show a "projectibility power". But this is just an illusion: all the green samples of jade cannot imply the law "Jade is green", because this is merely a conjunction of two other hypothetical laws –namely, "jadeite is green" and "nephrite is green"–.

Kim asserts that in this example, there is no such a thing that can be considered a *standard kind* (not Kim's expression) named "jade", at least not in the way the term kind is normally used: the term "jade" rather refers to 'a disjunction of two heterogeneous nomic kinds' (Kim 1992, p.12), a *disjunctive* kind. As a matter of fact, this is the reason why there is no legitimate inductive projection that can be drawn from a law concerning disjunctive kinds. The failure is clearly explained by the

author: if a single kind is used to refer to both nephrite and jadeite, all the samples that have ever been examined might be revealed to be samples of jadeite, rather than a mix of both minerals. As a consequence, the law cannot be extended to nephrite because there is no experience of this particular mineral. Accepting this law-like form would entail accepting that any law "X and Y are G" can be verified by "X is G" and, at the same time, this last law would imply "Y is G", independently of what Y is. Of course, such a procedure is dismissed by Kim as illegitimate and confirmative.

The reasoning that leads to local reductions is then straightforward. Jadeite and nephrite are considered distinct on the basis of their microphysical nature, that is to say, they show chemically different reactions (i.e. they are composed of different molecules), despite the fact they show similar macro-properties. Using this perspective, these two minerals can reasonably be considered two realizations of the kind "jade", but nobody seems to have problems with their reduction to two different micro-structures, characterised by different chemical features. Therefore, Kim's suggestion is to consider "pain" as a disjunctive kind and, consequently, to consider it a non-nomic class, characterised by a virtually endless set of nomologically possible realizations (animal species, artificial systems, alien forms, etc.).

Kim's position explicitly leads to the assumption that multiple realizability entails that mental kinds can be never considered "scientific" kinds because of their disjunctive feature. Nonetheless, it cannot be concluded from this that psychology is a pseudo-science, because it is still grounded on physical realizations: i.e. it concerns entities which are explainable in terms of physical laws and processes. The difference with another science such as, for instance, chemistry, is that psychology is only reducible within "fields" described by means of specific conditions: each law established within each field determines a local reduction and this is the only "field" where it can be applied validly. In the case of psychology, these conditions are (roughly, as it has been mentioned before) described by a species-specific structure.

Finally, the inevitable conclusion of the author's reasoning consists in claiming the multiple realizability of the whole psychology, as a direct consequence of the multiple realizability of each mental state:

> If physical realizations of psychological properties are a "wildly heterogeneous" and unsystematic lot, psychological theory itself must be

> realized by an equally heterogeneous and unsystematic lot of physical theories. (Kim 1992, p. 20)

Since each realization entails a different local reduction grounded on a different "specific local law", an immediate objection to this theory consists in stressing how easily we are forced to allow local reductions which are only valid within small groups of organisms, or maybe even a single organism within the same species. Kim uses two arguments to defend his hypothesis from this criticism: first, neurophysiological researchers seem to have good reason to think that organisms belonging to the same species show important similarities that make it plausible that they also have similar realization processes (or local laws). Secondly, the physiological-biological differences should be considered in the same way as we usually consider psychological differences: if we are able to focus on the regularities of the psychological states, the same strategy must be conceded to the analysis of the structures. The differences are not simply ignored, on the contrary: psychological differences are explained thanks to differences in physiology.

Incidentally, both these arguments can be considered the embryonic stage of Bechtel and Mundale's appeal to a "different grain of analysis", a thesis that the authors described few years later in their article 'Multiple realizability revisited' (1999). Interestingly, Kim uses them within the context established by multiple realizability, whilst I will describe in section five how Bechtel and Mundale developed and used the same arguments to reject the theory, starting from its premises.

Directly mentioned in Kim's article, Fodor wrote a strong criticism to this thesis in a follow-up article titled "Special Sciences: still autonomous after all these years" (1997). Kim's entire proposal is rejected by Fodor, who attacks the foundations of the argument: concerning the kind "jade", Fodor denies that this kind is an example of multiple realizations and questions about the reason why it is unprojectible. The pivotal reasoning in this criticism consists in considering the example of the jade as an example of sampling error, rather than an example of disjunction. Fodor states that should we discover that all our samples of green matter called "jade" are samples of jadeite, this finding would simply prove that we are grounding our inductive hypothesis on a biased population. A population of samples which is biased can be used neither to support a generalization, nor to deny any law: '*The sampling error*

means that the data are equivocal, not that the hypotheses are unprojectible.' (Fodor 1997, p 152. Italics in the original text).

Fodor concedes that jade can be used as a good example of disjunctive kind, but he stresses that this is not relevant for the debate: what functionalists argue is that there is a substantial difference between a property that is disjunctive and another that is multiply realized. According to the author, a properly named disjunctive property, for instance composed of "kind A" and "kind B", is characterised by the metaphysical existence of worlds which can have either A or B or both. In contrast, a property which is disjunctively realised relies on different bases in different worlds.

Fodor's point is then explicit: the reason why pain can be considered nomologically homogeneous depends precisely on the ability to distinguish between these two kinds of disjunctions, conceding the functional description of pain its supervenience towards the heterogeneity of the physical matters that are contingently realizing pain itself.

4. Challenging Multiple Realizability. Sporadically in the early nineties and more frequently in the last decade, the once unquestioned plausibility of the arguments in favour of the multiple realizations of mental states has been challenged in several articles. Suddenly, it has appeared that the whole idea of multiple realizability could be rejected because ill-grounded and vague in some of the terms used.

The first example of this kind of radical criticism against the theory is represented by Zangwill's short article "Variable Realization: Not Proved" (1992) in which the author briefly analyses three distinct arguments in support of the hypothesis of multiple realizability of mental states, pointing out what he thinks are their weak points.

The question raised in the article is clearly stated by the author himself: given the negative claim of multiple realizability (asserting the non existence of a cause-effect relation between type-mental and type-physical) and the consequent functionalist positive claim about a higher level of description which relies on the analysis of the functional states, the presence of functional isomorphic devices which show physical differences is a necessary consequence. However 'how do we show that such possible functionally equivalent systems would also be in the same mental

state, and thus that it is functional organization which is important and not physical constitution?' (Zangwill, 1992 p. 216).

It is important to stress that Zangwill considers functionalism as always asserting the identity between mental and functional states: however, as it was described in section two of this dissertation, this assumption can only be considered true if it is applied to Putnam's early version of multiple realizability theory and, as a consequence, to the early version of functionalism. As a matter of fact, Fodor's generalised version already reached a level of abstraction that allows overtaking these identities. Despite these considerations, Zangwill's question should still be considered as worth receiving an answer which is far from obvious.

The first thesis analysed is defined as 'the argument from imagination': in this dissertation it has been referred to as the "conceivability argument" and it can be summarised in the mental experiment that allows us to imagine an alien organism or a silicon based device capable of being in a human mental state. Zangwill establishes a parallel between this mental experiment and another one which has been used as an anti-functionalist argument: the latter is centred on the problem of 'qualia' (i.e. qualitative or perceived experiences). The problem is that we can imagine two different organisms functionally isomorphic or sharing a particular functional state in a specific time and in spite of this, we can still imagine these two organisms experiencing (or being in) a different *quale*. Zangwill states that the common functionalist answer to the qualia objection results in a boomerang against the theory itself: the counter objections are summarised by the author as the claim that what is conceivable shouldn't be necessarily considered possible or even real. Were it so, the fact itself that we can imagine such "conceivable situations" still would have to be shown (for instance, it seems that we imagine inverted qualia, when we actually do not).

Thus, the author points out how the argument against the "quale hypothesis" can be easily used against the mental experiment of the alien organism or the silicon based device able to be in a human mental state. Zangwill's analysis entails a divergent path that starts from the use of the argument from imagination. On one hand, accepting the functionalist counterargument implies a refutation of the appeal to the superimposition of what is conceivable over what is possible. On the other hand, if we want to accept the value of these mental experiments, we are committed to accept that

it is possible to appeal to a further, higher, level of abstraction that denies the uniqueness (and therefore the validity) of functional descriptions of mental states.

This second conclusion, which is not clearly stated in the article, but I think it is implied in the reasoning, is particularly useful for the purpose of this dissertation and the analysis of multiple realizability theory. The appeal to abstraction and a higher level of analysis, which are necessary for a functional state description, might lead to an infinite process towards further levels of abstraction. If the functional states are the descriptive superstructures of the organisms, the qualia can be conceived as a super-superstructure of description: the problem starts if we accept that this process can be replicated infinite times, continuously appealing to new structures characterised by a higher and higher level of abstraction.

The "conceptual argument" is the second analysed by Zangwill: he summarises it as the thesis that considers mental properties as essentially having a causal role in relation to other mental states and behaviours. Conceptually, the essence of mental properties does not rely on any physical matter but on its role and of course this assumption implies the multiple realizability of the causal roles (and consequently of mental states) independently of the contingent physical matter that realizes them. The author states that this assumption is implied by 'a description theory for the meaning of mental terms' (ibid., p.217), but it is perfectly possible to consider the role of the mental terms as a way to fix *a posteriori* their reference: in other words it is possible that philosophers are establishing how to define mental states after their causal role on other mental states and behaviours. If the second hypothesis is assumed true, the identification between mental states and causal roles becomes contingent whilst the identification between mental and physical states is again a legitimate hypothesis.

In conclusion, the author does not express a preference as to the two types of relation established between causal roles and mental states: the point is that the question once again cannot be considered trivial and any answer is controversial. As a result, Zangwill disputes that a theory claimed to be undisputable and so widely accepted may be grounded on the conceptual argument.

The third argument is named by the author as 'empirical' and it refers to the possibility of hard evidence that may clearly support the theory of multiple realizability. In his article, Zangwill states that currently there is no empirical evidence in favour of this theory, neither from an organism substantially different from the ones characterised by neural systems, nor from artificial systems such as

computers –whether composed by silicon microprocessors or cogs and wheels– which might have implemented forms of artificial intelligence that replicate the same mental states that characterise a generic living being.

The likelihood of a hypothesis cannot be considered evidence by itself: it needs at least some weak empirical evidence to rely on, otherwise it falls in the first category of arguments analysed, based on conceivability instead of empiric proofs.

Every known organism assumed to feel a mental state such as pain is characterised by a neural structure which shows many similarities across species from the point of view of its working 'mechanics' ruled by chemical and physical laws. Even if this argument does not imply that in the future a different form of organism (such as an alien) or an artificial device would be discovered able to realize any of the known mental states of living beings, it must be granted that such a system has not been found yet so that *de facto* there is no empirical evidence that can be used to the advantage of multiple realizability.

A few years later, Shapiro wrote 'Multiple Realizations' (2000) which shares with Zangwill's article a similar structure, but focuses on a different aspect of the theory. Incidentally, it is interesting to notice that whilst he was writing, Shapiro was not aware of anybody else's attempt to question the truth of the multiple realizability theory[7] (as he states in the premises of his article). A fact that testifies to the absolute rarity of these attempts of arguing against the general grounds of multiple realizability, until the beginning of the current decade.

Shapiro's main interest is in the substantial vagueness of one of the fundamental assumptions of the theory: he claims that, even among the supporters of the theory, there is little or no agreement concerning the conditions that must be satisfied in order to recognise without any doubt the presence of multiple realizations. The author states that nobody ever tried to establish any sort of condition, underestimating the whole problem of giving a definition about the circumstances to be satisfied. As a result, it is not possible to state clearly what kind of physical differences are necessary in order to conclude that two devices showing them and simultaneously realizing the same function can be properly said as displaying 'distinct realizations'.

[7] Apart from Zangwill's article, it appears that Shapiro was also unaware of Bechtel and Mundale's "Multiple Realizability Revisited" which was published in 1999 (the next section is entirely dedicated to their analysis).

Before describing which conditions should be taken into account, Shapiro briefly analyses what he considers the main arguments in support of the theory, partially overlapping the categories already described by Zangwill. Firstly, the author gets rid of the argument grounded on the existence of functionally isomorphic systems: according to him, this is an *a priori* argument about physical relations because it follows directly from the definition of functional isomorphism. As a consequence, it cannot be used as empirical evidence as it is sometimes stated. What is more, the concept of functionally isomorphic devices cannot be used to support the hypothesis that these two devices are realizations of the same kind (which is the author's way of referring to the two devices realising the same function). Shapiro is convinced that this mistaken identification between functional isomorphism and multiple realizability must be ascribed to the erroneous assumption that a system is identical to its 'description of sequences of relations among functional states' (Shapiro 2000, p. 639). To underline this gross mistake, the author uses the analogy of a mousetrap and its description or representation on paper: according to the author, these two elements are functionally isomorphic because they preserve functional relation between the states, as required by Putnam's definition of functional isomorphism. Despite their isomorphism, nobody would seriously consider the two systems as capable of performing the same function, accomplishing the same task or in any way performing distinct realizations of the same kind.

As a second, Shapiro deals with the likelihood argument. As it has been described in the previous sections, in 1967 Putnam claimed that, if we assume any psychological state to be identical across species (hypothesis already challenged when Shapiro writes his article, as it will be shown in the next section), there is a high probability that this state is realized by systems which are characterised by substantial physical differences. However, as Putnam already stated and Shapiro underlines, it is still possible that parallel evolution converges on the same structures: thus, this argument cannot be used as the definitive reasoning against reductionism. After all, nature provides several examples of highly improbable events that happen anyway[8].

[8] The most unlikely event I can imagine is represented by a casual combination of organic molecules generating an aggregate characterised by a helicoidal shape which starts replicating itself. Nonetheless this is what, approximately, happened around four billion years ago when the first forms of viruses started to populate our planet (of course I am not considering creationism and "God's will" as a possible theory to take into consideration).

Since the main arguments fail to support concretely the existence of multiply realized kinds, it is necessary to establish rules that can be applied in all the necessary circumstances and that can be used to tell the difference between proper and illusive multiple realizations. According to the author, 'To say that a kind is multiply realizable is to say that there are different ways to bring about the function that defines the kind' (ibid., p. 644). These differences must affect directly those properties which have a causal relevance in generating the function under analysis: thus, firstly, it is necessary to establish which function is going to be analysed in the device and, secondly, among the features that characterise the device, it is necessary to understand which are the few ones that are causally relevant in relation to the selected function.

According to this model, it is perfectly possible that two devices showing substantial physical differences may not exemplify yet multiple realizations of a function. The reason is quite simple: once the function has been established, we could change every physical feature that does not have a direct causal role in generating it and it would not matter. The example presented by Shapiro is centred on the mental experiment of replacing single neurons one by one with silicon-based chips: Zangwill would have categorised this example in the group of those committed to the problems related to the arguments from imagination, whilst Shapiro simply states that even if it were possible to realize these replacements, it could constitute a physical change which might have no causal relation with the main function of the brain (i.e. bringing forth the mind in Shapiro's conception). The argument paradoxically implies that the only way to be sure that the mind is multiply realized is by studying the brain in the first place, in order to be sure of the causal physical relations that are responsible for the mind itself. As a consequence, the multiple realizability theory, which was supposed to deny the validity of reductionism in the study of the mind, becomes, at this stage of Shapiro's modification of the theory, an important reason to push ahead a reductionist approach.

There is still another consequence to be drawn from the rules identified by Shapiro: differing in those features that are causally relevant, the devices which realize the function effectively become of two different kinds. As a consequence, a valid multiple realization would be identified by means of a rule that never allows to verify it: on one hand, if the differences between the devices do not affect causally relevant features, we are analysing the same kind realised by devices that must be

considered of the same kind, identical within the perspective of the function realised. On the other hand, if the differences affect causally relevant features, we are analysing two incomparable devices: if two devices have nothing in common but their functions, the only type of laws that can be drawn from their analysis is of the analytical kind. That is to say, such a condition simply allows to state 'the capacity in virtue of which a functional kind is the kind that it is' (Shapiro, 2000 p. 649): the author fairly adds that stating, for instance, that all different kinds of eyes can see, as you can draw from the very definition of what is analysed, does not constitute a significant knowledge improvement.

Let us consider again the usual conclusive table:

Absence of causally relevant differences:

$A = M$
$B = M$ $\qquad A = B$

Presence of causally relevant differences:

$A = M_1$ $\qquad A \neq B$ and
$B = M_2$ $\qquad$ no relation (but analytical) can be established between M_1 and M_2

The first case shows a perfect identity between the two elements in the equation: the systems (*A* and *B*) must be considered identical because, relatively to the function they are implementing, they do not differ in causally relevant features. The second table shows the opposite case: when the devices differ in their causal features relatively to the implementation of the analysed function, there cannot be established an empirical comparison between the two cases: the diversity among kinds in the first element of the identity makes every empirical statement impossible.

This is called by Shapiro the multiple realizability dilemma": if the author's hypotheses are right, once it has been established a function, whether two devices differ in causal relevant features or not, the theory shows to be completely incoherent because of the theoretical impossibility to come out even with a mental experiment that could support it.

Obviously, this reasoning has major consequences on the concept of the autonomy of the special sciences as Fodor formalised it. The focus must switch from the study of the laws characterising the functional kinds to the study of the diversity in

mechanisms categorised as having a causal role in generating[9] the function analysed. Using Shapiro's example, "being able to see" is an analytical statement that allows to establish the category named "eyes" so that the functional kind have the role to fix the science domain, facilitating the empiric investigation of those physical laws which are causally relevant.

In 2001 Rosenberg wrote 'Comments and criticism on Multiple Realization and the special science' which is the only critical assessment I am aware of, dealing with the problems arisen from Shapiro's thesis. In his article, the author analyses Shapiro's position about the plausibility of multiple realizability and the autonomy of the special sciences, drawing the attention essentially to two arguments. The first one appeals to the heterogeneous nature of the domains of special sciences: considering an evolutionary perspective, it is plausible to concede nature the ability to find more than one way to solve a problem. If compared to the similar argument already present in Putnam, Rosenberg's position grounds the contingency of reductionist relation in the dynamic feature of evolving systems: the environment continuously challenges the organisms to fit new conditions and all organisms represent themselves an ever-changing variable to one another. If we apply these rules to the special sciences domains, systems must compete against one another to fit their environment in a more rapid way if compared to the times needed by organisms to evolve in nature. As a consequence, the chances to have causally relevant changes in realizers of the same functional kinds are exponentially greater in number, making a reductionist perspective unlike to be effective if not in short periods of time. The law concerning the reduction becomes, in the author's words, 'historically bounded' or 'temporally true' (Rosenberg 2001, p. 368).

The second argument relies on the alternative Shapiro gives us concerning the autonomy of special sciences: according to Rosenberg, the first option consists in accepting that functional kinds have single realizations and, as a consequence, there can be no autonomy in special sciences (i.e. reductionist perspective). Otherwise, multiple realizability may be saved at the high price of denying the existence of laws (whether contingent or not) connecting functional kinds to their realizers and therefore special sciences are not empirical sciences (i.e. eliminativist perspective).

[9] In Shapiro's conception of the multiple realizability theory, the mechanisms cannot be considered responsible of *realizing* the function as it was established in the precedent versions of the theory.

The problem is that, in Rosenberg's analysis, multiple realizability is considered almost inevitable due to the way evolution works both in nature and in special science domains (chiefly referring to emotions, singular and collective behaviours). As a consequence, Shapiro's reasoning leads to a model that is committed to the eliminativist perspective and the attempt to save the existence of special sciences assigning them a taxonomic rule generating classes seems to confirm this way to interpret the given alternative: 'once their [of the special sciences] usefulness in taxonomizing explananda has been exhausted, they have no further function' (Rosenberg 2001, p. 373).

Shapiro clearly did try to avoid such a conclusion, and Rosenberg himself considers it highly unreasonable: the problem both the authors are well aware of consists in the fact that, once such a position has been embraced, it is unclear how, in the past decades, it has been possible that the models arisen within special sciences could show the explanatory powerfulness everybody tributes them.

5. Towards the supervenience of neuroscience. The main dispute the theory is facing at present is fought in the field of neuroscience: new empirical evidences and models of mind arise from this field and they can hardly be shown to be compatible with the theory here discussed.

A strong use of neuroscience findings against the conceivability of the multiple realizations of mental states has been recently attempted in Bechtel and Mundale's 'Multiple realizability revisited' (1999). The previous section has presented the first switch of attention in the analysis of the theory: rather than considering the classical analysis of the theoretical implications that can be derived from Putnam's hypothesis that functional kinds are multiply realized over physical kinds (i.e. Fodor's and J. Kim's choice), Zangwill and Shapiro have focused their arguments firstly on the conditions that must be considered necessary in order to verify the presence of multiple realizations and, secondly, on the role the functional states play in the relation established between them and the physical kinds.

Bechtel and Mundale represent a further change in the strategies of analysis of the theory showing a perspective which is reversed if compared to the last one: the authors focus their analysis on the physical part of the relation and, coherently, on the practical consequences the theory has on the paths of research in neuroscience. Their

objective is made explicit in the article: they want to deny the validity of the antireductionist conclusions implied by Fodor's generalised version of the theory presenting mainly two arguments.

First, they claim that it is undeniable that in the last decades a huge amount of information has been being acquired about brain structures of species other than human and their correlated psychological predicates. Even though this correlation does not support the old identity theory, these findings are consistently affecting the understanding of human structures and the related mental processes: it cannot be ignored that this fruitful path of research would have not been even opened, if the generalised version of the theory –and its hypothesised contingency in the relation between physical and functional kinds– had been taken for granted.

Secondly, it is supported the idea that the notion of a physical or chemical brain state is, quoting the authors' words, 'philosopher's fiction' (Bechtel and Mundale 1999, p. 177). According to the authors, such a problem with the reference to one of the two elements in the relation analysed so far, takes the main responsibility for the practical fallacies of both the identity theory and the multiple realizability theory.

Concerning the first of the two points, this empirical way to support the hypothesis for cross-species neural structures is known to be a weak defence of reductionism: findings which support across-species neural correlates for specific mental states would not prove that in the future a different form of device or organism might show to be able to implement one of those mental states thanks to a different structure. In a few words, even an endless series of proofs in favour of the across species neural identity would not make the identity necessary and would not establish a law-like relation between the two elements considered. Nevertheless, it must be reminded that the majority of the arguments in favour of multiple realizations do not rely on a rare chance to find two different physical kinds realizing the same mental state: they assert firmly the overwhelmingly high probability to find in nature several physical devices able to perform such realizations. As a matter of fact, Fodor's reasoning brings forth the idea that examples of such kind have already been found.

Nevertheless, neuroscience has changed dramatically since Fodor wrote about neural plasticity appealing to the physiological psychologist Karl Lashley: the hypothesis of multiple realizations over times has been grounded on a vision of the brain which has been meanwhile abandoned by science. Bechtel and Mundale draw in their article a detailed picture of the evolution of models within neurosciences,

reconstructing in particular the development of the researches oriented on the generation of neuroanatomical maps that could assign mental functions to specific regions, especially considering the cortex. Lashley's criticisms to these studies, which were centred both on specific human cortex regions and across species structures, were grounded on the many ambiguities that characterised at that time the identification of these regions. In the attempt to criticise Korbinian Brodmann's hypothesis of functionally "static" cortex, Lashley and George Clark took two spider monkeys (*ateles geoffroyi*) and, using the same techniques, they reached two significantly different neuroanatomical maps of the cortex of the two apes.

In spite of the results of experiments like the one just described, whose attempt was to show the inconsistency of such hypothesis, the studies for a more accurate map of the human cortex went on, reaching controversial but useful results. However, the key moment for these researches has been reached in 1990 [10] thanks to the introduction of a new tool of investigation: the functional magnetic resonance imaging (fMRI) which was added to other new techniques of tomography. The fMRI is a non invasive technique which is becoming as more important as more the improving technology allows fast and highly defined images: its main contribution is still today in the fact that fMRI is the only tool that allows the analysis and the record of the activation state of groups of neurons in human patients awake and aware. Combining these new tools of imaging with posttraumatic studies, the cortex has been recently divided into nearly fifty functional areas, each showing an even further division into a high number of cortical sub-regions.

It is undeniable that there are many ambiguities in the results coming from these studies: first, it is not easy to find a task that may result in the activation of a single area. Secondly, should such a task be found, it is highly unrealistic to consider a human being as capable of focusing all his or her attentive resources on this very single task, becoming suddenly oblivious of everything that does not concern it. Third, but most important, the controversy about the existence of species specific

[10] The first experiments that used magnetic resonance for the purpose to discover functional regions in central nervous systems have been described in: S. Ogawa, at al. 1990: "Oxygenation Sensitive Contrast in Magnetic Resonance Image of Rodent Brain at High Magnetic Fields", Magnetic Resonance in Medicine 14: 68-78. Thanks to the functional magnetic resonance, it is possible today to avoid the use of invasive electrodes and researchers are now able to study what "happens" (what regions are "active") in healthy human brains as well as in damaged ones, when the patient is asked to solve abstract or practical problems.

neural differences has been fed by these new findings and ways of investigations: neuroanatomical maps show unique features (only valid in the subject analysed) as well as the ones that can be applied to a wide variety of species (some regions of the cortex) and others that can even be generally applied across species (some neural structures in proto-brains share their functions with superior mammal neural systems). Above all, the already known phenomena of plasticity and degeneracy within a single subject have been widely confirmed by fMRI. It seems clear now that these differences may not be ascribed to the vagueness of experiments or the mistaken offsetting of background noise.

According to Bechtel and Mundale's reasoning, these phenomena can be easily explained appealing to the very definition problem that constitutes the original ground of that "philosopher's fiction" already cited, that is to say the erroneous idea that there is a scientific use of terms such as physical or chemical brain state. It is acknowledged by the authors that the closest neuroscientist concept refers neither to a specific activation of single neurons nor to the transmission of single or multiple impulses through synapses or nerves. Neuroscience is in fact familiar with the different idea of a "neural activity", which refers to all the variations in energy use that characterise the passing from a generic rest status to an energy consuming activity that can be located in a certain region or area of the brain.

Consequently, the authors point out that this philosopher's fiction misled in generating the series of interpretative fallacies so far described by means of the different "grain of analysis" that has been consequently assumed by the communities of philosophers and neuroscientists. In other words, the supposed existence of a psychological state such as hunger in octopuses and human beings (the authors have chosen intentionally an example already used by Putnam in his 1967 article) can be only considered correct if we associate it with a general feature, as it would happen with a vaguely described behaviour. If we accept that "hunger" is the "food-seeking" behaviour, then we are adopting a *coarse grain* in the psychological analysis and it seems just fair to use the same kind of rough analysis in relation to the neural structure responsible for the behaviour. Given this exceptionally coarse grain of analysis, the neuroscientists may conclude that there is a neural activity in a part of the neural system of the octopus, which is not incompatible with a correspondent activity in human nervous systems. If the mental states are conceived as different as

the behaviours they generate in the two organisms, the simple existence of a neural activity is itself enough to be accepted as a similarity.

For the same reason, if we prefer a fine grain method in the analysis of the precise neural structure responsible for the psychological state of hunger in octopuses and human beings, then it becomes clear that the two systems are completely different, but at the same time, they are also realizing two mental states which are at least as different as their neural correlates. It should be conceded that, even if we only consider the external behaviour linked with the psychological state of hunger, a food-seeking octopus does show neither the same sequence of actions, nor the same singular behaviours of a food-seeking human. The inference allowed is that, in this new and well-detailed context, the two systems are not realizing the same token-mental state and it is extremely hard to believe that these two mental states may even belong to the same type.

According to the authors, the fallacies have been then created by the attempt to match a coarse grain analysis for functional kinds, generating a class of generic states such as hunger, with an extremely fine grain analysis for physical kinds. Furthermore, this latter kind of analysis is so *extremely* fine, that it ended in the creation of a kind that cannot be considered but 'fictional', since it is not used in any of the sciences involved in these investigations: this is precisely the case of the activation of single cells or the identification of single impulses firing through specific synapses.

In his article 'Discussion: a defense of Bechtel and Mundale', Couch (2004) fairly points out that the argument for multiple realizations always requires both a difference in the physical states and, symmetrically, a coincidence in the psychological predicates. Bechtel and Mundale's appeal to the different grain of analysis makes the argument fall either in the first step (in the "coarse" case) or in the second one (in the "fine" one): it seems virtually impossible to verify it under any circumstances.

This new reductionist position is summarized in the following table[11]:

[11] I am using the symbol "≈" (approximately equal) to stress the difference with the identities in the precedent tables. In place of the identity between mental states and "C-nerve fibres activation" or firing neurons, there is now a correlation with the neural activity in a region. The area involved may be identified with good approximation but still it cannot be established a theoretical identity involving each single neuron which is part of the cortical area described to be responsible for the specific mental state analysed.

Fine grain case:

$A \approx M_1$
$B \approx M_2$ $\quad A \neq B$ and $M_1 \neq M_2$

Coarse grain case:

$A \approx M_1$
$B \approx M_2$ $\quad A \approx B$ and $M_1 \approx M_2$

The same reasoning may be used to give an account of the differences noticed in neuroanatomical mapping of cortices. It is not simply a way to stress similarities where others have observed differences: the main effort made by the two authors is directed towards a philosophical appraisal of the findings coming from neuroscience. The identification of brain areas based on their activity during a correspondent behaviour (or supposed inner psychological state) does allow looking for homologies within different brain structures and, as a matter of fact, it enables successful generalizations across species.

An interesting objection has been arisen in the article 'Testing Multiple Realizability: a Discussion of Bechtel and Mundale' by Sungsu Kim (2002). Even though he considers the thesis for multiple realizability not proven, the author focuses his attention to the necessary distinction between evolutionary concepts of homology and homoplasy; a distinction that he claims being perfectly appropriate for brain structures. On one hand, every similarity in the structure of different species which is based on their descent from a common ancestor defines an homology: a commonly used example shows the connection between bird wings and human arms which are characterised by the same structure (if we use the perspective of their bones structures) even if of course their use is now completely different. On the other hand, homoplasy is defined as a correspondence between structures or organs (or parts of organs) acquired as the result of parallel evolution: in this case, the typical philosophers' example is represented by the structure of camera eyes in humans and octopi.

In his article, S. Kim claims the homologies in the brain to be unsuitable evidence both against and in favour of multiple realizability theory. On the contrary, homoplasies associated with similar functions performed by two or more species, may help to determine whether it is possible for a certain psychological state to be realized

by different structures: for instance, bat wings show a completely different structure when compared to bird wings, despite the fact that they are associated to the same function –to fly–. If we consider these circumstances, the absence of homoplasy becomes a proof in favour of multiple realizations of the specified function, since it is *de facto* realized by different structures.

In conclusion, S. Kim suggests that the homoplasies may be used as the *litmus paper* of multiple realizations: on one hand, positive examples are the essential proofs against multiple realizability; on the other hand, negative examples are necessary in supporting the theory. S. Kim's remarks are then addressed against Bechtel and Mundale's researches: which he sees focused only on homologies across species which will not prove anything concerning the theory. The reason has already been mentioned at the beginning of this section and relies on the fact that even if the convergence of identical brain structure would be proven in any circumstances, this would still be a weak argument against multiple realizability allowing in a hypothetic future the discovery of a psychological predicate characteristic of a living being, realized by silicon-based devices or by other life forms (the "usual" aliens from other planets) not based on neural systems.

In his article, Couch (2004) replied to the objections raised by S. Kim: the author concedes that there would not be any informative content in a study that only focused on homologies, when these are realizing similar functions. Nonetheless, Fodor's generalised hypothesis is centred on the ability of a single structure to perform different tasks in different periods. Thus, even if there is certainly a difference in the study of a single structure and the study of how these structures evolved within brains across species, this research performed by neuroscientists and used as an argument by Bechtel and Mundale is still useful in order to belittle the likelihood of generalised theorists' predictions as characterised at least by great imprudence.

Aside from all the considerations about homologies and homoplasies, it seems fair to concede that neuroscience has proved so far to be fruitful for the studies of the mind, so that Bechtel and Mundale's questions in relation to Fodor's generalised theory deserve to be taken into account. In particular, the authors' concern is that the assumption of the predominance of virtual machine studies over studies centred on neural systems is blind in respect of those empirical findings that have been successfully using the hypothesis of across species constancy of the neural correlates,

implicitly denying the contingency in the relation between mental and physical. From a pragmatic point of view, Fodor's generalization proves to be poor and to constitute an obstacle to future achievements.

Finally, the two authors devotes little attention to the hypothesis of the development of a device able to realize psychological states but made out of components that do not share anything in common with the living organisms. In this case, if such a system were built, even the appeal to the coarse grain type of analysis would not save from conceding the existence of a proper multiple realization of mental states. The authors do not to consider this hypothesis likely to happen even in a remote future: they claim that such a result would probably be obtained with artificial machines far different from the present serial ones. Moreover, in spite of everything claimed by proponents of multiple realizability, there are widely accepted cases in physics and chemistry in which a particular set of properties is only realised by a single entity (the authors use the example of water functional properties, which can only be realized by the molecule H_2O): the psychological state may be one of these.

The authors' conclusions concerning the functionalist approach in general may come as a surprise. They state explicitly that their reasoning does not directly affect functionalist assumptions *in toto*: their challenge to multiple realizability does not undercut functionalism because the functional characterization of the mental states is still important in order to fix the identification of those kinds that neuroscience later attributes to brain regions.

> Moreover, as we have emphasized, it is the functional characterization (the contribution to behaviour) that guides the identification of brain areas. A thoroughgoing functionalism uses functional criteria to identify both psychological states and brain states and can survive even if we jettison the multiple realizability theory. (Bechtel and Mundale 1999, p. 204-205)

This conclusion is interestingly close to the one Shapiro will support a year later in the article already analysed in the precedent section, but it seems to me that the whole problem has been voluntarily ignored in order to downplay its importance. I will show in the next chapter that, starting from the same premises, other authors

coherently assume a completely different position in relation to the supervenience of the role of functions in brain mapping researches.

All things considered, it still seems that, methodologically and theoretically, the multiple realizability theory is not as “overwhelmingly likely” as it has been assumed since Putnam’s first formalization and it is gradually losing its strong attractiveness.

Chapter 2

6. Parallel processes: from functional states to m-functions. In my opinion, two problems have affected the multiple realizability theory during its whole evolution and both of them seem to have a major influence on the arguments and the reasoning used by both the supporters of the theory and those who try to demonstrate its fallacies. I define these two problems as follows:

1. The identification of the mind with the probabilistic automata is structurally inadequate to give a computational description of processes performed by parallel distributed systems.
2. The abstract use of the term "function", which characterises the generalised theory, leads to partial description of mental processes, depriving the theory of its efficacy and descriptive power.

These problems share the same origin and can be easily considered the two sides of the same coin: one of the main issues that contributed to their generation is represented by a slight change in the reference and in the concept of the Universal Turing Machine when these systems started to be used in the field of philosophy of mind. The definition of such virtual machines, which are pivotal for all the different versions of the multiple realizability theory and in general for the functionalist model of mind, was originally fixed by Turing's theory of computability. The problem is that, when contemporary philosophers deal with Universal Turing Machines, they are actually giving life to a different type of system characterised by different computational features. These features are usually implicit, lacking the mathematical definition that is typical of the proper Turing Machines, and this is the main reason why they are rarely being questioned.

To put it simple, from a computational point of view, the identification between the computation performed by a neural system and a Turing Machine (or a probabilistic automaton) is widely overestimated and, in addition, it is usually erroneously attributed to Turing himself. A list of misconceptions about Turing's famous article "On Computable Numbers, with an Application to the Entscheidungsproblem" (1936) is well described by Copeland's entry in the Stanford Encyclopaedia of Philosophy "The Church-Turing Thesis" (2002). In brief, Copeland addresses as a philosophical 'myth' the idea that in his article Turing may have

mathematically demonstrated how a *Universal Turing machine can compute any function performed by any other machine* with any architecture, given sufficient time and memory.

The fallacy that relies on this sentence can only be explained after giving an account of how two specific words are used within the computability theory entailing a reference different from the one usually attributed.

First, another good way to describe this false hypothesis is "*a Universal Turing machine can compute any set of instructions* [etc.]". Thus, the term "function" is used in this context as a synonym with "set of instructions" or "program" of a virtual machine. In the first chapter it has already been pointed out how several authors dealing with the problem of multiple realizations have different ways to use the term *function*. Its reference varies from the functional states in a virtual machine, to psychological states in a living being or to the tasks accomplished by an artificial device, but in none of these circumstances it has been used in place of the description of the state transitions, as it is used here. Yet, from a mathematical perspective, a function is a relation or an expression involving one or more variables: this is the reference that the term has when it is applied to the field of computation. As a consequence, a *function computed* by a Turing machine corresponds to its set of instructions rather than to one of its functional states because it describes the way information is processed by the system. Finally, considering the input of the machine, which is the information to be processed or –in a mathematical perspective– the domain of the variables computed, we can conclude that the functional state consists in a value assigned by the machine to a specific moment in its process, in relation to the assignation of values to the variables.

Since the analysed systems are virtual machine, a schematic result may be represented as follows:

$\{ x_1, x_2, x_3, \ldots x_n \} \rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow \ \ldots \rightarrow$ (final status)

The input assigns a value to each of the n variables $\{ x_1, x_2, x_3, \ldots x_n \}$, then the virtual machine computes these values as it is described by its set of instructions, reaching its first status (A). The new status gives life to a new series of processes that allow the machine to change again status in favour of the second one (B): the operation is replicated till the virtual machine reaches the final status described by the instructions in relation to the values assigned to the variables.

Questioning if the *function computed* by a system may coincide, under certain circumstances, with the *function* (i.e. the task) *accomplished* by the system itself, is a misconception of the terms used. Thus it is important to stress that the way the term function is used in the field of philosophy of mind differs substantially from its use in the field of mathematics (and consequently in the field of the computational theory). Thus, in order to avoid the misunderstandings that may be caused by this double reference, I will use the words "mathematical-function" (or m-function) to refer to the set of mathematical expressions that establishes the rules a system uses to process the information.

Finally, if we were only discussing processes in virtual machines, it would be possible to consider the use of the expression "set of instructions" or the term "program" in place of "m-function", since the first two have already been introduced in the first section of this dissertation. But the purpose of this reasoning is to demonstrate how some particular systems can process information in a way that is not computable by the virtual systems described by Turing: my proposal is then to use the m-functions to refer to the entire class of the hypothetic processes to whom it is possible to give a mathematical description. The use of a new expression becomes necessary because the precedent cited alternatives are conceptually bound with Turing machines and therefore they would mislead again into the problem that gave origin of this argument. That is to say, the fallacious belief of the identification of the class of processes which have a mathematical description, with the class of processes that can be calculated by Universal Turing machines.

As a matter of fact, it must be acknowledged that there are m-functions that cannot be computed by Universal Turing machines, even if they can still be computed by other systems.

What Turing did demonstrate is that a properly instructed Universal Turing machine can realize any m-function that can be computed by any Turing machine, but these m-functions must be characterised by a 'mechanical method' (simply *M* in Copeland's article) that must meet the following requirements:

> 1) M is set out in terms of a finite number of exact instructions (each instruction being expressed by means of a finite number of symbols); 2) M will, if carried out without error, produce the desired result in a finite

> number of steps; 3) M can (in practice or in principle) be carried out by a human being unaided by any machinery save paper and pencil; 4) M demands no insight or ingenuity on the part of the human being carrying it out. (Copeland 2002, first section in the web-page).

It may seem that the range of the m-functions computable by these virtual devices still remains vast; nonetheless, these four rules set a series of important restrictions, making the types of realizable m-functions decrease in number, especially if we compare this new range of possibilities with the order of infinite that characterised the first assumption about the omnipotence of the Universal Turing machine.

Therefore, it is clearly stated (requirements 1 and 2) that Turing machines are characterised by finiteness: finite number of steps for finite number of instructions (independently of the variable assignation) which are themselves defined by a finite number of symbols. As far as the other two requirements are concerned, these may be difficult to interpret in the proper way: Copeland exemplifies the idea appealing to the use of the truth table test for tautologies. The test must be conceived as applied to every m-function just in principle, despite the fact that it is easy to imagine why, in practice, there are not enough pencils and papers in the world to work out some complex m-functions which are computable by Turing machines.

At any rate, for the purpose of this dissertation, it should be sufficient to point out that the set of hypothetic m-functions realized by any Turing machine is *countable*, that is to say, it is characterised by the same order of infinite of the integers. Since the number of all the hypothetic computable m-functions is of a higher order of magnitude, it is said that it is *uncountable*[1]: hence, there is an infinite number of m-functions that cannot be realized by a Turing machine and therefore cannot be considered Turing-computable.

If the first cause of misunderstanding about the notion of computability relies on the use of the term function, the second one has been clearly pictured by Copeland as

[1] In mathematics it is possible to distinguish between orders of infinite: two sets composed by infinite elements can be compared and it can be established if one of the two is of a higher order, that is to say, it includes *more elements* than the other even if both of them count infinite ones. A simple example is given by the numbers: even though they are infinite, integers are clearly *less* (i.e. minor order) than real numbers. The latter class, in fact, comprises all integers as well as rational and irrational numbers.

generated by a false understanding of the statement that Turing machines can compute any *computable* m-function. This statement can only be considered true if we use properly the definition of "computable" given in the Church-Turing thesis: in brief, the thesis defines every m-function characterised by a mechanical method as effectively calculable and, as a consequence, computable. Clearly it is due to this definition of computability if the set of m-functions effectively computed by Turing machines coincide with the set of *all computable function*s. As Copeland underlines, the problems start when this notion of non-computability by Turing machines is confused with the one of the absence of a mathematical description.

This is the reason why I consider the use of the term "computability" in the way it is described by the Church-Turing thesis as easily misleading into the overlap of the set of computable functions with the decisively vaster set of mathematical functions. It has already been stated that the same Church-Turing thesis also allows the existence of mathematical functions which are not characterised by a mechanical method and consequently are not computable by a Turing machine, even though they can be computed by other systems using different forms of processes. As a consequence, I do not see a good reason to keep on using the Turing machine perspective in order to establish whether an m-function is (properly) computable or not, since it is clear that the same thesis allows the possibility for (properly) incomputable m-functions to be computable, once the perspective of a differently structured system is used.

Once again, in order to avoid future misunderstandings in this dissertation, I will always refer to the "Turing-computability", in place of the standard (and deceptive in my opinion) "computability" defined in the Church-Turing thesis. It must be borne in mind that this change in words does not change the mathematical reference: both expressions refer to the very set of m-functions.

In conclusion, the newly established class of m-functions is composed of those which are Turing-computable plus those which are "differently structured system"-computable. Incidentally, it is important to avoid another erroneous deduction which may be drawn by the absence of the explanatory power of a Turing machine for the latter type of m-functions. As it has been stated concerning the m-functions that these systems implement, processes performed by non-Turing systems are mathematically describable: every process implemented by a system has its own complete description the very moment the m-functions it realizes are formalized.

Before establishing whether or not the relation between neural systems and the mental is contingent, it appears now that, if we want to be adherent to Putnam's early formulation, there is a fundamental requirement to consider in order to have the chance to use properly a virtual machine as a description or simulation of what is implemented by the neural matter. Actually, if it were possible to show that the m-functions implemented by neural systems fail to meet at least one of the requirements that are necessary for the simulation with a virtual machine, then it could be concluded that this path leads to a dead end. Namely, it would be sufficient to show that the processes realized by these systems are not characterised by finiteness to have m-functions which are not characterised by the feature of being Turing-computable. Even assuming a functionalist approach, I think this reasoning might be sufficient to come to the conclusion that it is necessary to study the way parallel systems process their data, in order to set out a falsification path towards the hypothesis that relies on the famous identification of the mind with the virtual machine. What conclusion such a path would lead to cannot be established *a priori*.

It may be argued that even if we could find out that parallel systems do not realize Turing-computable m-functions, this finding by itself would not be enough to discard multiple realizability. A new hypothetical and more powerful virtual machine might be conceived: different from the known Turing machines, it might widen the range of computable functions realizable by virtual machines, overcoming some of, if not all, the weak points of the classic machines.

Nonetheless, it seems that a similar powerful virtual machine is unlikely to come and it is usually considered mathematically implausible[2]. But even if it were plausible, this objection would not lead far from a reductionist perspective: these new hypothetic systems should not simulate a generic new set of m-functions but those specific of the parallel distributed systems[3] so that once again, in order to be sure that

[2] The existence and the features of devices that may result to be able to implement such Turing-incomputable m-functions have been debated at least for five decades. An essential bibliography and a brief account of this debate can be found in section 2 of Copeland's entry in the Stanford Encyclopaedia (see references for details). As a matter of fact the probabilistic automata represent a virtual machine able to realize a wider set of m-functions, if compared to the Turing machines. In this chapter I will refer only to Turing machine for the convenience of the reasoning, but the criticism can be applied to the probabilistic automata as well (the set of m-functions realized is still *countable* and the m-functions themselves are characterised by the same features).

[3] In literature it is usually necessary to distinguish parallel systems characterised by local representations from those characterised by distributed representations. For the purpose of

the proper set of m-functions is part of the domain of these new machines, it would be necessary to know first what type of m-functions are implemented by parallel systems. Furthermore, such new devices would also have to be characterised by one of the essential features in the Universal Turing machines which are reprogrammable: input data can change the m-functions implemented (within the set of the Turing-computable m-functions) giving a good explanation of the process of learning from environment in living beings. The new hypothetic virtual machine must share this feature with the Universal Turing machines or it will be forced to simulate a fixed set of m-functions.

My claim is that, whether or not such new machines are necessary (and, consequently, whether or not the early hypothesis of multiple realizability is still plausible) can only be established after understanding what sort of m-functions are realised by neural networks. Thus, it is necessary to cope with the specific features that characterise a parallel distributed system and, taken for granted the definition of probabilistic automata, it is widely recognised that parallel distributed systems differ from the former serial machines concerning[4]:

1) Time and energy requirements.

2) Memory operation (storing and recalling information).

3) Input information encoding.

4) Autonomous development of categories.

5) Reaction to unknown stimulation (confabulation).

6) Robustness (ability to resist to physical damage).

7) Degeneracy (ability of structurally different elements of a system to perform the same function or yield the same output).

Biological systems based on neural structures require specific amount of time and energies in order to activate their systems. A lack of the latter may modify substantially the computational processes performed (affecting the neural processes in

---

this dissertation it is not necessary to investigate the former systems since the latter are closer to biological neural networks. This reasoning aims at showing what kind of computations are implemented by natural systems, in order to falsify the assumption that these m-functions can be realised by anything else but parallel distributed neural systems; as a consequence, unless specified differently in the text, every time I refer to parallel systems, I will then refer to parallel distributed structures.

[4] I give here a brief account of the main distinctive features that characterize parallel distributed systems: my first source in this task is supplied by the two volumes "*Parallel Distributed Processing: Explorations in the Microstructure of Cognition*" (1986) by McClelland, Rumelhart *et al*. (see references for details).

the nodes of the network) by the system independently of both the awareness and the perception of such a lack by the organism: in other words, the m-functions implemented may change independently of the input data, in consequence of a difference in the amount of the available resources. Figuratively, the system can be reprogrammed by inside mechanisms as well as by outside information.

Concerning time requirements, if the processes are suddenly interrupted due to a lack of time, these systems are still able to give an output even if it will probably differ from the one the system would have reached having the correct amount of time. Nonetheless, it is acknowledged that, once the device has started processing its input information, there is no datum that may result in a system failure (i.e. in a situation characterised by no output) because the output units of the system always show a pattern corresponding to a vector during the whole data processing. For the same reason physical damages of the network, or any other action that results in altering the normal execution of the processes described by the m-functions, will not result in system failures even if they will probably affect the output vector of the system.

The way information is stored and recalled is another fundamental feature that distinguishes the m-functions implemented by these systems from those implemented by serial machines. It is known that biological neural systems are able to distribute information between several units of the network at the same time. When information is stored, it does not become a copy of the original set of stimulations: it is rather modified depending on the precedent information stored. Similarly, when information is recalled, it may be modified again or it can be only partially recalled (e.g. few distinctive features of a complex set of stimuli).

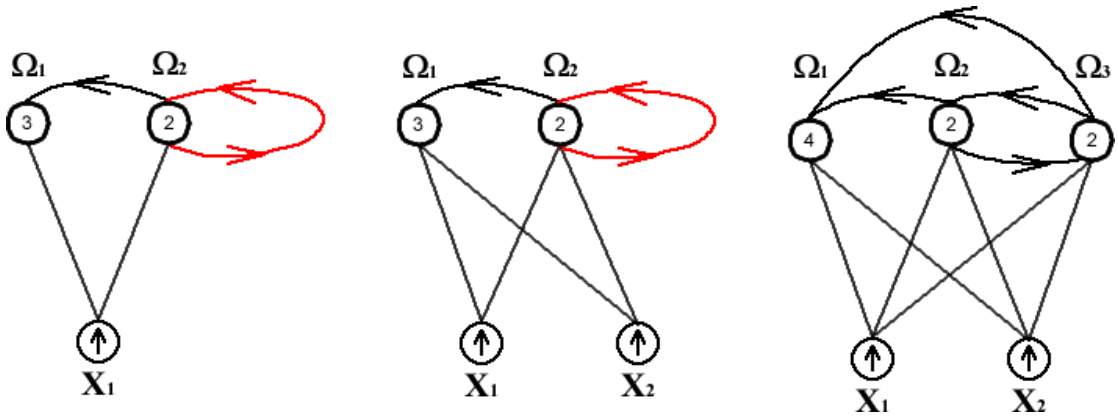
In order to clarify these particular features, it is probably useful to give an example: a few months ago I was asked to design a neural architecture that could check the correctness of a string of symbols, given a simple grammar and a vocabulary only composed of two symbols (for the purpose of the argument, let us say “1” and “2”). The grammar consisted in a single rule which stated that the device had to accept strings composed of no more than a single unit of the second symbol and any number of the first symbol: the symbol dispositions had to be considered irrelevant and, finally, any other string had to be rejected as incorrect.

In other words, the devices I am going to present here have to be able to analyse each string composed of the two elements in the vocabulary, in order to classify them

by means of an association to an output for acceptable strings and a different output for unacceptable ones. For instance:

12112	Not acceptable
121111	Acceptable
1111111112	Acceptable
...	...

There is no established limit to the number of symbols per string, and as a consequence the parallel system cannot rely on local representations. In other words, a neural system checking every single position of each string with a different receptor cannot be used: there is no alternative to a distributed system.



The three networks here drawn represent the scheme of three parallel systems able to accomplish the given task: in particular, the first scheme represents the simplest possible network able to solve this particular function, whilst the others show just two of the infinite possible variants of the first structure. Other architectures, characterised by a different conception and a more complex structure may be easily conceived: for the purpose of this reasoning, these three little networks exemplify the fact that infinite parallel structures may accomplish this specific task.

Thus, the first network is characterised by a single receptor X_1 (the circle with an arrow inside), two nodes or neurons which are the computational units Ω_1 and Ω_2

(the circles characterised by a number inside) and four links or synapses[5]. Since the device is asked to process a virtually endless string, I decided that the single receptor will process a single position in the string per segment of time (e.g. the first example given in the table, 12112, is decomposed in five inputs, the first signal is 1, the second is 2, then the third and the fourth is 1 again and finally 2 to close the string).

When the first symbol reaches the receptor, it is encoded as a number and then it is propagated unaltered through the synapses till it reaches the two computational units (the neurons) where it is processed separately. The neurons compute the signals and react in a way that is determined by the specific mathematical function that characterises each of them: one of the main features of these m-functions in the examples here proposed is represented by the threshold of the neuron. In order to simplify the argument, I here assume that these computational functions are always constituted by simple inequalities: as a consequence, if the signals reaching a neuron overcome its threshold (the value is expressed by the number pictured in the circles that represent the neurons), the neuron becomes "active" and it propagates a single signal (value=1) to all the outgoing links. Finally, when the neuron marked as Ω_1 is activated, the device classifies the analysed string as incorrect.

Concerning the structure, there is a final remark about the links in the neural network: it is necessary to underline that these connectionist devices usually show weights associated to these links, that is to say, a value is assigned to each of them and it strengthens or weakens the propagated signals. Once again, in order to simplify the description, these weights are all assumed to be equal to 1 (i.e. the signals are always propagated unaltered).

Let us consider again the first string in the given table (12112): the first position input is propagated by the receptor, it reaches the neurons, but it does not activate any of them. On the contrary, the second input unit activates the neuron Ω_2 which propagates the signal towards itself (by means of the red circuit) and the other neuron Ω_1. The third and the fourth inputs in the string do not change the conditions of the activation of the system: the second neuron is still active due to the signals coming from both the circuit and the receptor whilst the first neuron is still deactivated.

[5] In the pictures I assume that the signals go from bottom to top, unless an arrow specifies differently.

Finally, the fifth position of the string shows again the symbol "2": if compared with the first time this symbol has occurred in the string (i.e. the second position), now the neuron Ω_2 is already activated so that the neuron Ω_1 receives the signals coming from both the string and the network (black arrow-link propagates the signal coming from the activation of Ω_2 to Ω_1). As a consequence, even if the external stimulus is not different from the one already received in the past, this time Ω_1 is activated, giving the correct output that classifies the string as "not acceptable".

The m-functions realized by this system consist of two sets of inequalities:

t_0	t_1		t_0	t_1
$\Omega_2 = 0 \rightarrow$	**$x < 2$** $\Omega_1 = 0$ $\Omega_2 = 0$			
	$2 \leq x < 3$ $\Omega_1 = 0$ $\Omega_2 = 1$ **$x \geq 3$** $\Omega_1 = 1$ $\Omega_2 = 1$		$\Omega_2 = 1 \rightarrow$	**$x < 1$** $\Omega_1 = 0$ $\Omega_2 = 0$ **$1 \leq x < 2$** $\Omega_1 = 0$ $\Omega_2 = 1$ **$x \geq 2$** $\Omega_1 = 1$ $\Omega_2 = 1$

The table shows the two sets of inequalities in the columns t_1: the series of symbols ($\Omega_2 = 0/1 \rightarrow$) stands for "the deactivation/activation of neuron Ω_2 in the time t_0 implies the set of inequalities in the time t_1".

It is important to notice that the circuit (the red arrow link) in the architecture allows the input information to change the set of m-functions realized by the device (i.e. the way it reacts to its input). The signal propagated through the circuit keeps the neuron Ω_2 activated as long as the value assigned to the variable x (the input) overcomes or it is equal to 1; at the same time, the signal propagated by this neuron has the effect to change the "disposition" of the neuron Ω_1 to be activated by an external stimulus. Consequently, it would not be a mistake to claim that this single link grants the system a memory-like mechanism, allowing it to recognise the symbol "2" when it is showed for the second time.

The three neural networks previously drawn have been claimed to accomplish the same task: it is time then to show if it is possible that they are multiply realizing the same m-functions (i.e. the parallel system equivalent to the set of instructions in serial systems) or, what is more important as far as this dissertation is concerned, the same functional state (i.e. their equivalent to parallel systems, which are still to be defined).

The second neural network shows a single difference if compared with the first one: the architecture has been changed. The number of receptors has been increased of a single unit and, consequently, there are also two more synapses that link the new receptor to both the neurons. Such a difference implies that the input is now composed of a two-dimensional vector, that is to say, the "world" this new device has access to is composed of couples of numbers, instead of single numbers.

An input that may count on a vector with two dimensions allows a wide variety of encoding solutions to the problem of feeding the system with the usual string of two symbols[6]. Each of these solutions leads to different assignations of the other variables in the same set of m-functions, in order to make the system able to accomplish the given task (reaching the objective of classifying the strings). However even if every possible variable is kept constant, the main difference in encoding the signal (i.e. the vector dimension) has already affected significantly the m-functions implemented by the system, making them completely different from the ones presented to describe the first neural network.

As a matter of fact, in the given example, the network still sticks to the rule to use simplified versions of these parallel systems –fixed synapses weights (which propagate unaltered signals), thresholds (the same displayed in the first device) and computational functions in neurons (again simple inequalities)– but every alteration in its architecture results in realizing m-functions belonging to different sets. Furthermore, some of these structural changes make the m-functions so different that they can no longer be compared with one another: this is the case of a change in the number of receptors as in the example here analysed.

[6] For instance, the first easiest choice might be to mirror the encoding system already used for the first device, limiting the duty to propagate all the input signals to one of the two receptors. Another possibility would be to make one of the two receptors react to both symbols whilst the other one only reacts to the second. One more hypothesis would be to dedicate each receptor to only one of the two symbols in the vocabulary so that they are activated alternatively, depending on the symbol showed in the correspondent position.

The reason why the m-functions that describe the processes in the second network cannot be compared with those implemented by the first network relies on the impossibility to compare two vectors with two different dimensions or (which is the same in this case) two m-functions relying on a different number of variables. Namely, the vector $\{x\}$ is not comparable with the vector $\{x_1, x_2\}$: they properly refer to two different "worlds".

Therefore it does not matter which particular way may be chosen to encode the input signal: if, for instance, the first receptor propagated the signals mirroring the way the single receptor in the first device propagates its signals, and if, at the same time, the second receptor did not propagate any signal at all, still the input signals of the two networks would be completely different. As long the "silent" receptor is connected to at least a node of the network the input code pertaining to the two devices will generate for usual string (12112) the following vectors:

{ (1); (2); (1); (1); (2) }

referring to the first network and

{ (1,0); (2,0); (1,0); (1,0); (2,0) }

referring to the second network.

Finally, this is the set of m-functions generated by the second device:

t_0	t_1		t_0	t_1
$\Omega_2 = 0 \rightarrow$	$\mathbf{x_1 + x_2 < 2}$ $\Omega_1 = 0$ $\Omega_2 = 0$			
	$\mathbf{2 \leq x_1 + x_2 < 3}$ $\Omega_1 = 0$ $\Omega_2 = 1$ $\mathbf{x_1 + x_2 \geq 3}$ $\Omega_1 = 1$ $\Omega_2 = 1$		$\Omega_2 = 1 \rightarrow$	$\mathbf{x_1 + x_2 < 1}$ $\Omega_1 = 0$ $\Omega_2 = 0$ $\mathbf{1 \leq x_1 + x_2 < 2}$ $\Omega_1 = 0$ $\Omega_2 = 1$ $\mathbf{x_1 + x_2 \geq 2}$ $\Omega_1 = 1$ $\Omega_2 = 1$

The third example leads to a similar conclusion: this time the vector of input signals is unchanged if compared to the second device, but the first layer units covert

now the incoming signal in a three dimensional vector, which takes the place of the precedent two dimensional vectors realized by the other two structures. A neuron replaces the former red circuit and the threshold of the neuron Ω_1 changes for the first time (the new threshold, 4, is written within the black circle), so that it will be activated only if it receives the correct input signal from the receptors and an activation signal from the other neurons.

I think it is no longer necessary to show once more the new set of m-functions realized by this third network: the vector conversion differs from the former ones and, consequently, the description of the processes that make it possible must be different too.

Summing up, a neural network realizes a sheaf of sets of m-functions[7] defined by its architecture and by the computation performed by each single node of the network. The values assigned to the other variables, the main one represented by synapses weights, fix the constants for any specific set of m-functions within this sheaf. Every modification in the architecture of the network or in the processes of the single nodes leads to a system that could or could not solve a specific given task. For instance, considering only the usual simple inequalities such as the processes performed by the neurons, a neural network necessarily requires at least one of these architectural features in order to accomplish the known task[8]:

1) A circuit.

2) A link between nodes characterised by the same distance[9] from receptor units.

3) A feedback link that can propagate the signal coming from a neuron located at a higher distance to one located at a lower distance from receptors.

---

[7] For instance: the equation (ax + by = k) describes a sheaf of straight lines. If we fix the constants (in this case: a, b, k) attributing them a value, the result is the equation of a single straight line (e.g. 2x + 3y = 1). A set of straight lines describes more than one equation in which the constants have been fixed, when they are combined in single or multiple systems.

[8] This is a well known problem in the field of neural computation. The logical operator XOR is an example of this kind often cited in literature: it is known that there is no way to simulate this function with a single layer neural network (i.e. the architecture must consist of at least two layers). Many manuals analyse this problem: see, for instance, Stuart Russell and Peter Norvig 1995: 'Artificial Intelligence: A Modern Approach' Prentice Hall (chapter 19, section 3) or Jeffrey L. Elman et al. 1996: 'Rethinking Innateness – A Connectionist Perspective on Development' (p. 354)

[9] In the graph theory (which is the mathematical theory of network), the "distance" $d_G(u, v)$ between two nodes u and v in a graph G is the shortest path (i.e. the shortest number of nodes) that separates them.

When we work with simple connectionist models, the sheaf of m-functions implemented is easily describable as previously presented. But even if the systems show a higher order of complexity (such as those proper of biological networks), it is possible to have an idea of the sheaf of m-functions determined by the architecture, especially considering that, though extremely complex, single neurons compute their electrochemical signals in a way that can be described with sets of m-functions.

Finally, it should be clear by now that the three systems here pictured realize different m-functions, require different amount of energy and time to perform the same task and differ in the way the information is encoded or stored, in the categories developed and in their resistance to physical damages. In a few words, if we want to use Putnam's identification with the Turing machine, they rely on a completely different set of instructions. But this conclusion seems to deny the main concept implied by the multiple realizability: how is it possible that contingency characterises the relation between mental and physical kinds, if all these variables entail a substantial difference in the m-functions realized?

Anyhow, the problem that has given origin to this section is that it is not clear if any serial system such as a Universal Turing machine may o may not simulate the processes performed by parallel system: it seems to me that such a hypothesis is becoming less and less attractive. It can be argued that we could focus on the very task we are interested in, but this choice would lead to a partial description of the three systems: a complete description (which was in my opinion implied in Putnam's early version of the multiple realizability theory) may only rely on the simulation of the whole set of m-functions.

Still, the conceivability of a Turing machine able to replicate the m-functions of a biological system may be considered sufficient as the required evidence in support of multiple realizability. It has already been stated that the only way to find out if this mental experiment is true, is of the empirical kind. That is to say, it is necessary to verify if the set of "parallel system"-computable m-functions is a subset of the set of Turing-computable m-functions. At present we can only try to argue if such a hypothesis is as plausible as Putnam argued forty years ago.

I consider three main reasons to doubt that it will be ever possible to proof such a hypothesis, especially if we consider the distributed processes that characterise biological neural network:

1) In order to simulate m-functions of parallel systems with a virtual machine, we should take into account the valid argument Block and Fodor (1972) use to discard the FSIT theory[10]. According to the authors it is not possible to give a good explanation of simultaneous mental states using a single virtual machine. Nonetheless, it is also impossible to give a description of the unicity of the set of m-functions performed by parallel systems using several probabilistic automata operating in parallel, unless we do not use them to simulate the processes in each single node of the network. This solution would anyway lead to a dead end: taken for granted the supervenience of the architecture in the generation of the sheaf of sets of m-functions per parallel system, the complete description of the system cannot rely on the simple gathering of the neural processes.

2) It has already been mentioned that serial systems operate differently from parallel ones concerning the autonomous development of categories and (due to the same reasons) the reaction to unknown stimulation. These two features relate to the particular nature of parallel systems and I think they grant them the ability to work in the domain of infiniteness: a domain that the mechanical method cannot describe and that consequently leads to Turing-incomputable m-functions. It is known that parallel systems can deal with sets of infinite symbols as an input, being able to apply their processes to unknown stimuli as well as to known ones. It is true that external data can reprogram a Universal Turing machine to make it deal with new symbols: once the input has changed the set of instructions, the device can apply its rules to the once unknown data. Nonetheless, such a process does not simulate what happens in parallel systems which do not necessitate to change their m-functions in order to deal with the new input: the system automatically assigns a category to every input and it deals with that depending on the category. Once again, in order to simulate all the processes, the virtual machine is asked to deal with a potentially infinite set of symbols and, consequently, with an infinite set of instructions (clashing with the second requirement of the mechanical method), otherwise there will always be a difference between m-functions of the simulated system and those of the simulative system.

3) It has been mentioned that deficits in the amount of energy may have a decisive influence on the m-functions the biological neural networks realize. The three

[10] See section 2.

structures presented as an example should also give an idea of another way to modify the m-functions, independently of the input of the system. Physical alterations (mainly structural damages or other changes in the architecture and chemical or electrical interference in electrochemical synapses or in the metabolic state of the neurons) directly modify the way the information is processed by the system, but cannot be considered as part of the input. A simulation with a Universal Turing machine cannot replicate these phenomena, despite the fact that they are mostly frequent in all living beings based on neural systems. To establish a comparison, physical alterations in serial devices produce system failures most of the time. Fodor used the argument of plasticity and degeneracy to propose his generalised version of the theory[11], but it seems to me that this argument can be of good use also against the virtual machine hypothesis, at least until these systems will be able to realize m-functions only reprogrammable by means of the input.

In conclusion, my thesis is that even if we can all agree that parallel systems realize some sort of m-functions that can be mathematically described, it does not follow that these m-functions can be multiply realized. Consequently, it cannot be drawn the conclusion that the relation established between the m-functions and the neural systems is contingent. Actually, the more science gives us tools to investigate these parallel distributed systems, the more it seems that the processes they implement are necessitated by the physical matter and are characterised by a series of unique features.

In my opinion, Turing machines do represent an interesting example that helps us to understand what the relation between a device and the m-functions it realizes is, but it seems to me "overwhelmingly implausible" that these machines may ever be able to generate a complete simulation of the m-functions realized by parallel systems.

The reasoning so far described leads to a controversial conception of the mental states: it has already been established the approximate identity in the following proportion:

set of instructions : *Turing machine* ≈ *m-functions* : *system mathematically describable*

It may be argued that this proportion implies the following new one:

[11] See section 2.

functional state : *Turing machine* ≈ *assignation of values to all variables in the m-functions* : *system mathematically describable*

In the set of parallel systems (which is a subset of the systems mathematically describable), this proportion would imply that a particular kind of "activation pattern" would take the place of the third term. Though different from what Bechtel and Mundale defined in their article as "philosopher's fiction"[12], this would be anyway a completely theoretical object: a sort of photography of the entire structure, considering the whole network, the activation and metabolic status of all neurons and disposition of every synapse to propagate their signals. Consequently, every change in any of the variable involved, would generate a different set of instructions as well as a different "mental state", a conclusion that seems to be highly counterintuitive.

The alternative is also interestingly challenging: if the simulation with a virtual machine is not possible, maybe the second proportion is simply nonsensical. After all it may be necessary to take into consideration the hypothesis that there is not an equivalent in parallel systems for the functional state in Turing machines.

Incidentally, it is to notice that these two alternatives coincide with the two "genuine options" J. Kim grants to a physicalist approach[13]: if the second one is clearly a form of eliminativism, the first is can be conceived as a hyper-local reductionism, that establishes bridge laws for each singular structure over all its realizable "states".

7. Pursuing a complete description (or: concrete problems of abstract functions). The precedent section constantly refers to the lack of plausibility to realize a *complete* simulation of all the processes implemented by parallel systems, due to their uniqueness. Assuming that the simulation has to focus on the problem "which device can compute what kind (sheaves) of m-functions", it seems to me that there is no alternative left to an empirical investigation that might look for an answer to the question arisen by the proposed thesis: viz, do probabilistic automata realize m-functions implemented by parallel distributed systems?

Nonetheless, the immediate objection is: why should we accept this assumption? Or, in other words, why should we focus only on *complete* simulations?

[12] See section 5.
[13] See section 3.

After all, there are many other strategies that seem to lead to good explanatory models of the mind. Furthermore, these strategies are also perfectly compatible with the multiple realizability theory and, consequently, the theory can support them with its strong appeal.

I have briefly mentioned in the precedent section that Putnam's hypothesis concerning the isolation (and description) of single mental states was influenced by the identification of the information processes in organisms with the ones in probabilistic automata. It is hard to imagine that, at the time he wrote 'Psychological predicates', the author meant anything different from what he clearly stated, for instance, in the following two passages:

> All organisms capable of feeling pain are probabilistic automata. (Putnam 1967, p. 31).

> [...] identification of psychological states with functional states means that the laws of psychology can be derived from statements of the form "such-and-such organisms have such-and-such Descriptions" together with the identification statements [...] (Putnam 1967, p. 33).

It is not stated that they process information in a *similar* way: it is rather assumed that the probabilistic automata do simulate all the information processes implemented by any organism capable of feeling pain (i.e. in Putnam's conception, all animal species). Furthermore, the Description (capital letter in the original quoted text) of the organism relies on the functional description of the probabilistic automaton that is simulating its processes. Therefore everything the theory implies, such as the truth value of the identification between mental and functional states, is a direct consequence of the truth value of these assumptions concerning the equivalence of information processes and their description in virtual machines and organisms.

At any rate, the theory soon evolved abandoning the hypothesis of the identity between organisms and a single perfect description[14]: Putnam himself changed his mind in a series of articles published after his famous 1967 'Psychological

[14] Block and Fodor's arguments against FSIT: see section 2.

Predicates', claiming the multiple realizability of the mental states over the functional states[15]. In his entry in the Stanford Encyclopaedia, John Bickle writes:

> Notice that this argument for functionalism [i.e. Block and Fodor's argument against any type identity] is explicitly non-deductive, in contrast to the deductive (and valid) nature of Putnam's original argument against identity theories. It is important to keep the anti-identity theory argument separate from the pro-functionalism argument, as some criticisms of multiple realizability may be telling against one but irrelevant against the other. (Bickle 2006, third section in the web page: see references for details)

An analysis of the generalised version of multiple realizability becomes now necessary because it seems to be immune from the criticism already arisen against the early version of the theory, due to its peculiar differences. This new version changes the object of the multiple realizations, leading to the "function" mentioned as the second "side of the coin" at the beginning of the precedent section. This new function is still conceived as similar to the functional state in a virtual machine and therefore it is still separable from the entire set of processes realized by the system, but it characterised by this feature on the basis of a different reasoning. In my opinion, this is the cause of a new series of problems that affects specifically the generalised version of the multiple realizability theory.

It seems reasonable to gather all the paths that lead to preserve this particular feature of the function into three major categories even if they are related by the strong influence they exert to one another. The first choice is represented by a general concern about abstraction that I have already mentioned in the first chapter and that it is ascribable to Block and Fodor's 1972 article (the argument against the FSIT) and to Fodor's 1974 formalization of the generalized the theory.

The second alternative is grounded on the analogy between organisms and artificial devices: the idea is that it is possible to identify a single "main task" within a device and that the "function" has a similar role in biological (or in general information processing) systems. These type of functions can be either considered

[15] Hilary Putnam 1988: *Representation and Reality*. Cambridge, MA: MIT Press

self-evident or established a priori: the second variant is compatible with Shapiro's conception of a taxonomic value for the functions and it is directly used in relation to biological systems.

The final choice consists in claiming that a partial description may be sufficiently informative or characterised by a sufficiently elevated truth value: such a position usually aims at defending this type of investigation playing on its practical value.

Concerning the first alternative, aiming at a more abstract idea of the object of the multiple realizations should not be considered an operation that can come without a price. Once the identity between the functional state and the mental state is abandoned, if the pivotal term used in the theory lacks a substitutive clear definition, it will eventually affect with its vagueness the whole new version of the theory.

A major strength of Putnam's hypothesis relies on the simplicity of the argument:

1) Probabilistic automata can realize the same functional states, independently of the matter they are made of.

2) More than one organism is able to be in the same mental state X.

3) Organisms which are able to be in a mental state X are probabilistic automata.

4) Organisms can be in the mental state X, independently of the matter they are made of.

A probabilistic automaton has a unique best description of its processes coinciding with the formalization of its set of instructions by means of the adequate mathematical functions. Denying the existence of a best description for organisms entails denying the third point in the argument.

Therefore, it seems that the identity between the mind and the virtual machine becomes an explanatory analogy in the generalised version. It is no longer necessary for an organism to be a probabilistic automaton, but to *work like* this virtual machine, processing information in a similar way and, consequently, allowing multiple realizations.

If the theory were only grounded on its claimed plausibility, it would become a series of a priori statements as Zangwill hypothesised for the notion of functional isomorphism[16]. Thus, Fodor makes the theory avoid such a fate supporting the

[16] See section 4.

generalised theory with empirical findings as it has been described in the second section of this dissertation. But the identity between the functional state and the mental state is no longer granted, so that it can be questioned what the reference of the "mental state" exactly is, within this new version of the theory.

It may seem a paradox, but, not counting Putnam's first definition based on the identification, it is hard to find even a single definition in any of the articles dealing with the theory. The mental state is simply considered self-evident and characterised by features that can be established a priori: paraphrasing a judgement Paul Churchland (1981) expresses about the Folk Psychology theory (FP) in 'Eliminative Materialism and the Propositional Attitudes', the mental state is characterised by an abstract nature, whose inexplicability in terms of physical constitution is directly caused by the very abstractness of its nature: it is self-referential.

In the same article the author gives a detailed account of the reason why FP should be discarded: in brief, he asserts that the theory 'is a stagnant or degenerating research program and has been for millennia' (Churchland 1981, p. 75). What is more important for the purpose of this argumentation, when the author deals with the particular abstract nature of the functionalism in relation to the hypothesis of multiple realizations, he states that the theory expects the empirical findings to fit perfectly the a priori organization the theory imposes. In other words, functionalists use a method that subverts the standard scientific methodology: instead of grounding explanatory models on empirical findings, this theory builds the evidence to confirm itself. According to Churchland, the pivotal elements of the theory are self-confirmative, rather than self-evident.

The author draws this conclusion after the analysis of the examples often used in the reasoning in favour of multiple realizations: the articles usually refer to mousetraps, corkscrews, carburettors, arithmetical calculator, valves and many other artefacts built to accomplish a specific predetermined task. Churchland's criticism of this method of analysis is clear:

> Plainly, if FP is construed on these models, as regularly it is, the question of its empirical integrity is unlikely ever to pose itself, let alone receive a critical answer. (Churchland 1981, p. 78)

In addition, it seems to me that, if an abstract and all-embracing function becomes the object of multiple realizations, then the entire theory becomes as abstract as it is vague, losing its explanatory power. As Churchland points out, the theory inevitably becomes also unquestionable due to the elusiveness of its explananda.

The author fairly considers the examples based on devices as questionable: they usually lack in representing both the dynamic nature of biological systems and the complexity of information processing system. The problem vividly reveals itself when these devices are used to pursue the second strategy included in the list, which focuses on the supposed identification of a main task in a device.

There are at least three different ways to establish the function of a device and each of these ways may lead to a different conclusion. In order of importance: the purpose the device has been built for, its capacities (all the conceivable ways it can be used) and its primary uses (which may differ from those uses which were not "meant" when the object was created).

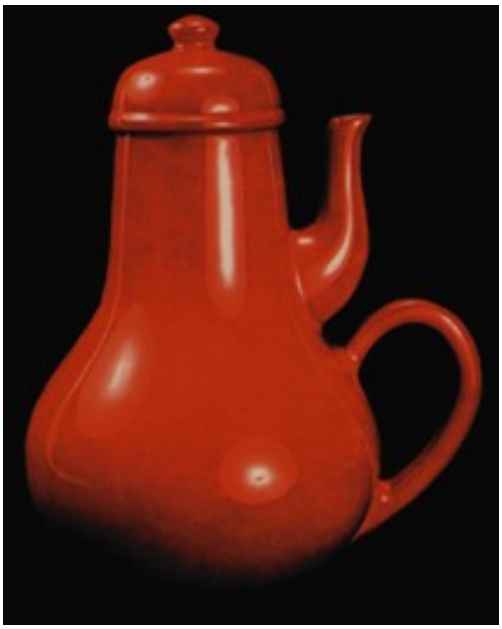
This problem is not new, but it has been probably underestimated even by those, like Churchland, who wanted to belittle the usefulness of these analogies with the artefacts. Shapiro for instance, rapidly addressed it[17]: concerning the way to define a function on the basis of 'purposes, capacities, contributions', he concludes that 'it may be possible that any of a variety of occupants can fill the role' (Shapiro 2000, p. 643).

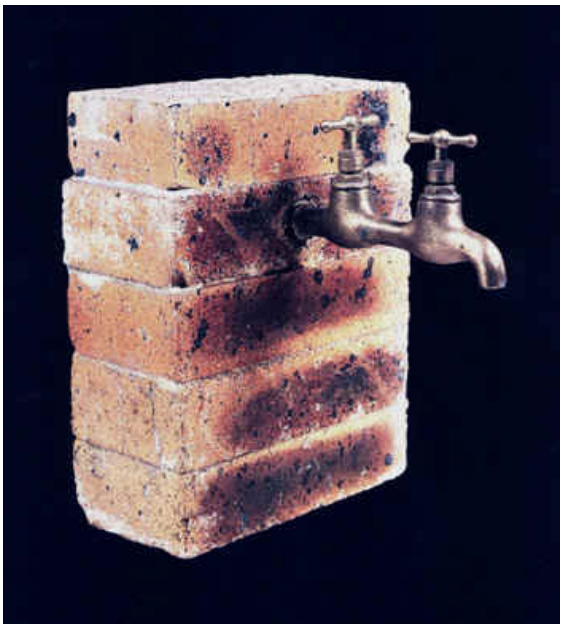
In order to make the reasoning evident, let us consider the following devices realised by Jacques Carelman and presented in his famous 'Catalogue d'objets introuvables' (1969): the "Masochist's coffee pot" (on the left) and the "Safety tap" (on the right). The former is probably perfectly described by its name, whilst the latter is described by the author: 'if you forget to close a tap, the other will prevent the waste of water'[18] (Carelman 1969, p.48).

An interesting question would obviously be "what is the function of these two devices?": in other words, we want to know what the purpose each device has been built for is, what their capacities, or their primary uses are.

[17] See also section 4 of the present dissertation.

[18] The book was first published in French and thanks to its success it was later on translated in other languages: I here report a personal translation from the 1978 Italian edition. See references for details.





In the first case, the definition of a function depends on the person who is describing it: a masochist will have a different idea about the *masochist's coffeepot* if compared to the one of a sadist or to the one of somebody who likes coffee and dislikes pain. Concerning the *safety tap*, it is stated that this device has the function to prevent the waste of water, but a quick analysis will easily reveal that the declared purpose of the author is nonsensical.

As a matter of fact, the function of these "unobtainable objects" as conceived by their author, was probably to stress in a peculiar way the importance and the principles of a correct design even in the simplest everyday objects. Nonetheless, if we do not know anything about the author or about ergonomic theories, once we are asked to do it, we will try to look for some practical use for these objects, whether or not the author of the device conceived the function we imagine. In addition, it is likely that, being obscure their "main function", each user will come out with a different idea about how to use the artefact and its capacities.

The second device allows stressing a further problem with this strategy: each tap accomplishes a different task depending on the condition of the other. If the first tap (the closest to the wall) is open, the second has the function to prevent or to allow the water to run out of the pipe. But this tap is deprived of its function the very moment the first tap is closed: things can get even worse if both taps are closed so that none of them, considered separately, has effectively any function at all. Once again, these arguments are not new: the investigation of the relation between cause and effect is

full of intriguing examples directed to challenge the idea of the existence of simple causes[19].

In conclusion, without a proper definition and left to the autonomy of the observer, the function of a device is under the authority of a point of view and a method of description. If this position is accepted, it entails that the function has become a chimera, even vaguer than the one generated by the aim at abstraction. Of course it can be objected that these objects are quite peculiar and have been chosen appositely because they can easily be negative examples for the strategy of the "main task". It seems to me that such a criticism against the choices of these two devices would have the side effect to strengthen Churchland's position against the choices of valves, carburettors and all the usual positive examples.

The use of simple devices as analogies for information processes in the organisms do not seem to lead to a fruitful path: if it is applied to selected devices, it starts a confirmative procedure; on the contrary, if it is applied to generic devices, it happens that some of them would become negative examples discrediting the hypothesis of the existence of a "main task" to associate to the function.

Despite all these consideration about the use of devices, it must be noticed that Carelman's objects, though odd, do not show the complexity of a biological system and it is legitimate to assume that this reasoning centred on the selection of a single function becomes even more complex if the analysis focuses on biological systems and mental processes. So that, if it is not plausible to establish the "main task" of a mental state in a way that can be considered undisputable or straightforward, it has been mentioned at the beginning of this section that a variant to this so-called second strategy consists in attributing a taxonomic value to the functions. That is to say, it may be established, as proposed by Shapiro, that the reference of the functions must be fixed a priori in order to become a guideline for empirical researches in brain mapping.

An interesting objection to this strategy relies on the possibility that our categories concerning the mental kinds are deceiving the physicalist approaches to the mental, once again generating confirmative procedures. This problem is being

[19] The tap example reminds me an example about two snipers often used in the theory of the cause. Considering two snipers shooting at the same time a deathly bullet against a target, if we apply a simple counterfactual strategy analysing the two events separately, none of them is effectively responsible for the death of the target. The action requires a complete description in order to be well explained.

investigated in the field of cognitive neuroscience: an example of this reasoning is Timothy Bussey and Lisa Saksida's (2005) 'Object memory and perception in the medial temporal lobe: an alternative approach'. Their findings come from the use of the fMRI and they have been used by the authors to overturn the usual conception of the normative role of the functions, re-establishing the supervenience of the physical:

> We would like to question the prevailing programme of trying to map psychological constructs such as 'perception', 'semantics', 'categorisation', and other notions onto anatomical modules in the brain. Evolution did not design the brain according to psychological categories that we have just recently invented, although neuroscience would be much easier had she done so. Instead of restricting ourselves to this way of thinking, we suggest attempting to understand the functions of brain regions in terms of what computations they perform, and what representations they contain. (Bussey and Saksida, 2005 p.736)

The position here supported is certainly controversial, but the idea that the taxonomic value of the functions may be considered an obstacle rather than a powerful tool, cannot be discarded without an accurate analysis. In particular, considering the empirical grounds of this thesis, for the time being the theory can be restricted within some specific fields, limiting its impact (the authors' studies are focused on the phenomena of perception and memory). Nonetheless, from the point of view of the computational theory this approach is extremely tempting. As a matter of fact, the last part of the quoted passage is particularly important for the purpose of this dissertation because it suggests the supervenience of the computation performed towards the –abstract– functional level of analysis. Using the terminology fixed so far, the article states that, in order to understand the processes parallel systems realize, it is necessary to formalize their implemented m-functions.

Finally, the last two steps in the present reasoning should lead to claim the weakness of the descriptions that come from the use of "function" in the generalised theory. First, it is important to remember that an analogy which is not characterised by an informative nature should be dropped. Secondly, the hypothesis of the isolation of

single functions in the particular systems that implement mental processes is ill-grounded.

Concerning the first, we can claim, with Rosenberg (2001) that an insect and a human being are just two of the numerous examples of multiple realizability of the function "life". These living beings share the same feature of being alive and are at the same time characterised by two completely different body structures. Nonetheless, using Bechtel and Mundale's (1999) terminology, their multiple realization of life is only similar in a coarse analysis: if we use a fine analysis the systems show their differences, but the function they realise can no longer be considered just the rough "life". Likewise, using Shapiro's (2000) theorization[20], it can be argued that this is not a proper multiple realization because it is applied to systems which do not differ in their "causally relevant properties" (i.e. they are both based on a helicoidal DNA or RNA molecular system[21]). In both cases the hypothesised multiple realizability is caused by both a generic definition of function and a mistaken interpretation of what a multiple realization should be. The hypothesis can be accepted as an example of multiple realizability, but in this case it does not result in a significant knowledge improvement or in a good description of the device.

An adage says: "there is more than one way to skin a cat". It is probably not controversial to concede that even if this statement asserts the multiple realizability of the function "to skin a cat", the description of reality it gives is not at all informative. It should be considered entailing neither an interesting use of the theory, nor a useful reference for the term function.

All paths here described are based on the idea that it is possible to select a single –best– function to assign to a mental state, isolating it from the rest of the system without altering its established function. This idea is supported by the analogy between the mind and the serial machine: going back to the exemplification of the state transition in these types of machines (A → B → C etc.), it is clear that each state can be isolated from the entire series of processes at no cost. The operation is also

[20] See section 4.

[21] Rosenberg uses the example of the prion as a molecule with hereditary function not composed of nucleic acid (i.e. without genetic material). Nonetheless this can hardly be considered as an example of living being and it is uneasy to show that this sub-microscopic protein can also have that hereditary function granted by the author. Once again the definition of the function realised grants or prevents the identification of multiple realization: this problem has been already addressed at the beginning of this section.

reversible: the set of instructions can be rebuilt putting in the right order each isolated state, so that the addition of the parts results in a description of the whole system.

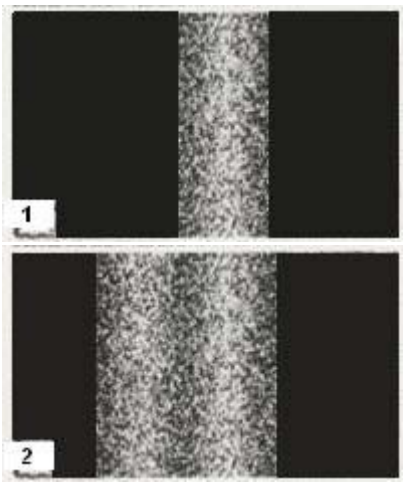
Due to the fact that the identity of the processes between serial systems and organisms (as hypothesised in the early theory) has been discarded and considering that an analogy has a weaker descriptive value than an identity, it is not granted that the isolation in a parallel system can be carried on without generating problems of any kind.

As a matter of fact, from a mathematical perspective, it is not possible to isolate a single m-function within a description set, preserving at the same time the original computational value. Each m-function in the set is generated by a computational unit (a neuron) and it is organised in the set depending on the architecture of the neural network, so that the isolation of a single m-function entails the isolation of the part of the network that is realizing it. The architecture of the system analysed is consequently different from the original one and so is the computation the new system processes. Every selection of particular m-functions in the set leads to the same conclusion: the very moment a singular or a group of m-functions is selected within the original set, the selection generates a different system and consequently the analysis has changed its focus. From the point of view of the computation performed, the only possible description of a parallel distributed system is the complete one.

The very idea of the partial descriptions as a general strategy is not free from theoretical problems. For instance, the *double-slit experiment* is famous in physics to show the dual nature of atomic and sub-atomic particles (i.e. they have a mass, occupy space and respond to forces but they also show a wave-like behaviour), entailing the dual nature of the matter itself.

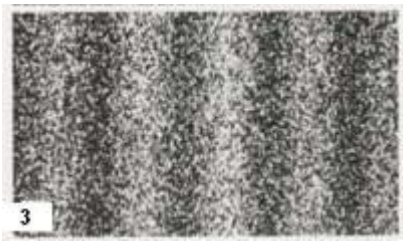
The explanation of the experiment requires the understanding of quantum physics and it is quite complex from a mathematical perspective, but for the purpose of this argumentation it is sufficient to describe the visible part of the experiment and what can be deduced from its observable results. Thus, let us suppose to have two screens, one in front of the other and the first one characterised by a slit, then let us randomly shoot small objects at the first screen. Some of these will go through the slit of the first screen, hitting the second screen: if they leave a mark where they hit, they will generate a pattern on the second screen, a bend whose dimensions will be similar to the ones of the slit of the first screen, with the greatest intensity directly in line with

the slit and rapidly diminishing towards the borders (image 1: each time an object reaches the black screen, it leaves a white dot as a mark).

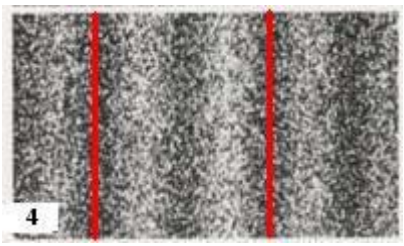


If we add a second parallel slit of the same dimensions and shape in the first screen and keep on shooting small objects at it, we will see a second bend on the second screen which will duplicate the features of the first one (image 2).

Let us imagine now to shoot atomic particles at the screen, say electrons: these particles are still characterised by a mass, so it is presumable that they will leave a track on the second screen that is similar to the one left by the "small objects" just described. In the first case, with a single slit in the first screen, the assumption is correct and electrons do behave like any other object characterised by mass. But once the second slit has been opened, we have no longer two, but many bends: this is the typical pattern generated by wave interferences, so that it can be concluded (roughly speaking) that particles behave like small objects and like waves at the same time (image 3).



Obviously, this is not the right context to go into details of quantum physics: what I am concerned of is that our focus of attention can easily give us a false idea of the phenomenon we are observing. If, for instance, we focus on specific parts of the second screen, trying to verify our first hypothesis (namely that particles behave mechanically like any other object in nature), we can easily locate the two bends we were looking for, being convinced that our assumption is right (image 4: two bends within the two red lines).



This way of dealing with the problem simplifies our understanding of the phenomenon: in the case of the interference pattern generated by particles, it is not possible to know, for instance, through which slit each particle goes, before it hits the second screen, leaving its mark. On the contrary, once we have selected the part of the second screen we are "mostly

interested" in, it is easy to explain the motion of each particle simply knowing where it hits the second screen.

Nonetheless, if our description of a phenomenon relies mainly on what we decide we want to focus on, the result is a description that leads to fallacies in the explanation model it produces. Sometimes a partial description is the only accessible description we can reach, but if it is only stirred by the choice of focus of analysis, then we are just using the wrong tool. The double slit experiment should exemplify the latter type of description: once we select the part of the second screen which is useful to verify our hypothesis, we manipulate the accessible data and what is a wrong explanation of a quantum phenomenon becomes perfectly plausible. Actually, such a selection would make the whole quantum phenomenon disappear in favour of the traditional mechanics.

If it is accessible, a complete description of a phenomenon is to be preferred if compared to a partial one. In the case of the double slit experiment, the mathematical description of the motion of the particles shot at the screens is tricky indeed[22], but it is the only one which gives us an explanation of the interference pattern visible on the second screen at the end of the experiment, once we decide to analyse it as a whole.

I think parallel systems show a similar complexity: we can focus on a single part of the processes implemented by the system, but in this case there is the chance that the selection is modifying the processes analysed. In particular there is the risk that our hypotheses might be confirmed because of the categories used in order to select the data to analyse. A complete description can avoid these obstacles because it is directed to the interpretation of the processes occurring in the system as a whole: in the case of information processing systems (as it is the case for parallel distributed networks) the m-functions represent this complete description. Interestingly, this kind of description is close to what Putnam aimed at when he conceived the functional state level of analysis: his intention was to have an explanation of mental processes that could be independent from the contingent occurrence of the visible behaviour and from the assumptions of the identity theory. His picture of the whole information processing relied on the description of the state transitions and the identification of

[22] The mathematical description relies on a series of equations which describe each particle going through both slits, none of them, just the first slit and just the second: it is as if it becomes a wave when it reaches the first screen and then it becomes matter again before reaching the second one.

mental with functional kinds: the strategy here proposed reaches the same objectives (independence from the observable behaviour and from the concepts of the identity theorists) relying on the mathematical description of the m-functions and challenging the value of the usual definition of mental state.

In conclusion, I consider multiple realizability as a powerful tool if restrained within specific borders. Anytime the processes realized by a particular system are inaccessible, the only way to attempt an analysis consists in assuming that another system, whose processes are accessible, is realizing some of the processes of the first inaccessible system.

The analysis is then narrowed to a part of the whole set of processes of the accessible system: as a consequence, the new aimed description is partial and it is referred to this last system rather than to the original one. In other words, an analogy has been stipulated.

Therefore my claim is that when multiple realizability is applied to neural systems, it is useful if conceived as an incomplete description of phenomena: a similar constraint does not entail to discard the theory as a whole because there are still cases in which there is no or little access to complete descriptions. Nevertheless, if a complete description is accessible, it must be preferred to the partial one achieved by means of the study of a similar system. If a process becomes suddenly accessible (as it has been happening in the past few years in the field of the mind studies), new descriptions will be formalized thanks to this change: on this new ground, new explanatory theories might be built, showing substantial divergence if compared with the ones formerly inferred on the ground of the indirect and partial analysis.

Whilst attempting to defend multiple realizations from Shapiro's arguments, Rosenberg states:

> In the philosophy of psychology, the multiple realizability thesis is a hypothesis advanced to explain the absence of discoverable psychophysical laws in a way compatible with physicalism. (Rosenberg 2001, p. 369).

It seems today that we are moving towards the finding of these laws: should this happen by means of neural computation, as it has been suggested, the multiple

realizability tool will see the fields it has been applied so far restrained, in favour of the new mathematical descriptions.

Conclusion

The use of the term *realization* is not metaphysically neutral: in the past forty years it has implied a position in support of the supervenience of the "functional level" of analysis. The thesis I support in this dissertation is clearly in defence of reductionism, still the massive use of this term seems to me that does not invalidate or confuse the reasoning.

My question relies on the use of this verb: when I ask "what m-functions do parallel systems realize?" I am assuming that the analogy between the mind and a virtual machine is at least an interesting way to focus on the study of information processes that are realized by the biological neural structure. Despite this disposition concerning the analogy, my point is that the question can only have an answer after an empirical investigation and, for the time being, all the evidence we have make the identification of the mind with any serial device extremely implausible.

Therefore, if it is true (as mathematics show) that processes in the brain depend on many variables, the most important being the architecture and the single unit computation, the consequence that I draw from the assumption that "*brain is realizing information processes*" is that reductionism is an inevitable path. This is the only path that allows understanding qualitatively and quantitatively the variables mentioned and, therefore, this is the only way to find out the sheaves of m-functions implemented by these specific systems.

There are many advantages in pursuing the use of mathematical descriptions as a tool to understand mind processes. The m-functions represent a complete description of the way every possible signal is computed by the system: they are not influenced by the presence of a specific stimulus or a combination of stimuli, neither they rely on the analysis of visible behaviours or other forms of output. As it was in origin conceived by Putnam concerning the set of instructions of a probabilistic automaton, the m-functions describe every possible process in the parallel system in each of its layers (input and output units included).

What is more, these descriptions can give an explanation of processes that are caused by external physical or chemical interference, such as electro stimulation or inhibition of regions of the nervous system, injuries that cause physical alteration of the structure, chemical interference in the metabolism of the cells or in the synaptic transmission. Just to mention the first example, there are surprising findings

concerning the micro stimulations of particular areas in the human cortex which generate sudden emotions in the patient: as I have tried to argue, these phenomena can hardly be explained with the classical use of the environment-reprogrammed Turing machine analogy.

I think that the main difficulty in *buying* this thesis does not come from my assumption about the empirical need of verifying the kind of m-functions performed by parallel systems. The problem is that, as I just mentioned at the end of the sixth section, the fate of the "mental state" is uncertain.

It has been argued that the conceptual possibility of isolating a single mental state, within the whole system of processes, has been influenced by the analogy with the Turing machine. My intention in the last section has been to show what reasons can be used to claim the inconsistency of this conceptual possibility, at least in the simple way it is usually taken for granted.

Therefore, if we exclude the eliminativist position, the following proportion may be used as the starting point for a new analysis:

functional state : *Turing machine* ≈ *assignation of values to all variables in the m-functions* : *system mathematically describable*

Biological neural networks are dynamical information processing systems, and consequently this reductionist perspective brings forth the concept of a "theoretical object" (or, as I have formerly defined it, a "photography" of the whole structure) that seems to be characterised by an unavoidable incoherence. This element can be considered sufficiently far from our perception of mental states to let us discard the whole thesis, on the basis of its implications.

I think this would not be a good reasoning: an analogy with the field of analysis in mathematics[1] should help in this case. A sheaf of straight lines can be studied both independently of the assignations of values to its constants and after the partial or complete assignation of the same values; the variables also contribute to locate specific parts or single points on the line analysed. As a consequence, it is perfectly plausible to imagine rules that can be applied in general to a biological parallel system (e.g. the computation performed by a single neuron is almost the same in every organism showing a central or distributed neural system: this is the assignation of value to a constant), others that are species specific (the macro structure of the neural

[1] The part of mathematics concerned with the theory of (mathematical) functions and the use of limits, continuity, and the operations of calculus.

network shows its similarities) and finally those rules which are single-structure specific and vary within a single organism depending on its natural development, experience and accidents. The use of the fine and coarse grain of analysis, as described by Bechtel and Mundale (fifth section), should make it possible to generate a way to relate "mental states" (or their equivalent hypothesised theoretical object) to the variances here described across species or within the singular organism.

In other words, this use of the mathematical descriptions does not lead to a hyper local reductionism: the single events in the flow of continuous processes of the system are still comparable within the same species with an acceptable fine grain of analysis.

Nonetheless, if it is applied an extremely fine grain of analysis, this reductionist hypothesis entails that even a single organism would never experience exactly the same mental state. The problem is then how to reconcile the "m-function" approach with our personal experience about the coherence of mental states: when we recollect a precedent feeling of, for instance, hunger it does not seem different from the present one, despite the fact that the structure in the brain has meanwhile changed in its micro features.

There are two hypotheses that have influenced my perspective concerning this problem. Gerald Edelman has formalized the first in a series of writings during the past two decades: his idea of the "remembered present" consists in assuming that a change in the structure of the system causes a change in the way memories are recollected and experiences are perceived because they are both implemented by the structure itself. The illusion of coherence is just the consequence of the change of the tool used to compare past and present experiences.

The second hypothesis has partially been mentioned in this dissertation and relies on Churchland's position against folk psychology: our perception about the coherence of mental states during our life may be as affected by fallacies as, for instance, our perception of motion of the sun during the day.

Symmetrically, going back to Putnam's starting point, the hypothesis of coherence of a single mental state across species, or even in alien species and computers, needs to be proved empirically or to be sustained by an adequate (complete) description of the processes that realize it. Otherwise, there is no reason to doubt that it may be inconsistent with scientific findings, despite its asserted plausibility.

Essential References

Bechtel, William and Jennifer Mundale 1999: 'Multiple Realizability Revisited: Linking Cognitive and Neural States' Philosophy of Science, 66: 175-207

Block, Ned and Jerry Fodor 1972: 'What Psychological States Are Not' Philosophical Review, 81: 159-181.

Block, Ned 1997: 'Anti-Reductionism Slaps Back' Noûs, Vol. 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World: 107-132.

Bickle, John (last modified 27 July 2006), copyright 2006 by John Bickle. *Multiple Realizability*. In *Stanford Encyclopaedia of Philosophy*. [Online]. Available: http://plato.stanford.edu/entries/multiple-realizability/ [15 February 2007]

Bussey, Timothy J. and Lisa M. Saksida 2005: 'Object memory and perception in the medial temporal lobe: an alternative approach' Current Opinion in Neurobiology, 15: 730-737

Carelman, Jacques 1969: *Catalogue d'objets introuvables*, (read in the italian edition, 1978) Milano, Mazzotta editore.

Churchland, Paul M. 1981: 'Eliminative Materialism and the Propositional Attitudes' The Journal of Philosophy, 78: 67-90.

Copeland, Jack B. (last modified 19 August 2002), copyright 1997 by Jack B. Copeland. *The Church-Turing Thesis*. In *Stanford Encyclopaedia of Philosophy*. [Online]. Available: http://plato.stanford.edu/entries/church-turing/ [13 June 2007]

Couch, Mark B. 2004: 'Discussion: A Defense of Bechtel and Mundale' Philosophy of Science, 71: 198-204

Dayan, Peter and L.F. Abbott 2001: *Theoretical Neuroscience – Computational and Mathematical Modeling of Neural Systems*, Cambridge, Mass., The MIT Press.

Fodor, Jerry 1974: 'Special Sciences (or: on the disunity of science as a working hypothesis)' Synthese, 28: 97-115

Fodor, Jerry 1997: 'Special Sciences: Still Autonomous After All These Years' *Noûs*, Vol. 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World: 149-163.

Kim, Jaegwon 1989: 'The Myth of Nonreductive Materialism' Proceedings and Addresses of the American Philosophical Association, Vol. 63: 31-47.

Kim, Jaegwon 1992: 'Multiple Realization and the Metaphysics of Reduction' Philosophy and Phenomenological Research, 52: 1-26.

Kim, Jaegwon 1997: 'The Mind-Body Problem: Taking Stock After Forty Years' *Noûs*, Vol. 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World. Philosophy and Phenomenological Research, 52: 185-207.

Kim, Sungsu 2002: 'Testing Multiple Realizability: a Discussion of Bechtel and Mundale' Philosophy of Science, 69: 606-610

Kripke, Saul 1972: *Naming and Necessity*, Blackwell Publishers, 1981 edition

McClamrock, Ron 1994: 'Kim on Multiple Realizability and Causal Types' Analysis, 54: 248-252.

McClelland, James L., David E. Rumelhart *et al*. 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition – Foundations*, vol. 1, Cambridge, Mass., The MIT Press

McClelland, James L., David E. Rumelhart *et al*. 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition – Psychological and Biological Models*, vol. 2, Cambridge, Mass., The MIT Press

Putnam, Hilary 1967: 'The Nature of Mental States', in Mind and Cognition - An Anthology, edited by William G. Lycan, 2nd edition: Blackwell, pp. 27-34.

Putnam, Hilary 1975: *Mind, Language and Reality. Philosophical Papers*, vol. 2. Cambridge, Mass., Cambridge University Press.

Rosenberg, Alex 2001: 'On Multiple Realization and the Special Sciences' The Journal of Philosophy, 98: 365-373.

Shapiro, Lawrence 2000: 'Multiple Realization' The Journal of Philosophy, 97: 635-654

Sober, Eliott 1999: 'The Multiple Realizability Argument against Reductionism' Philosophy of Science, 66: 542-564.

Zangwill, Nick 1992: 'Variable Realization: Not Proved' The Philosophical Quarterly, 42: 214-219.