Multiple realizations of the mental states: hunting for plausible chimeras

Vincenzo G. Fiore

1 The framework of multiple realizability

The key elements characterising the functionalist approach to mind studies are commonly identified (e.g. see [5]) with claims concerning:

1. The cognitive creatures' essential feature (they are all computational systems);
2. The object of the research in the fields of cognitive psychology and artificial intelligence (abstract functional states and novel physical realizations for these states respectively);
3. The irreducibility and consequently the autonomy of special sciences;
4. The inefficacy of the empirical research on the neural structure, because of the merely contingent relation established between the neural structure and the functional states it realizes.

The objective of this paper is to support a reductionist perspective in mind studies, disputing the soundness of the claims 3 and 4 in particular. Therefore, since it is easy to concede that the theory of multiple realizability of mental states plays the role of the hub, binding all the four claims one another, this paper aims at showing the weaknesses of the grounds on which the theory has been built.

The Multiple Realizability Theory (MRT) has been first formalized in the late sixties by Hilary Putnam in a famous series of papers (for a collection see [12]). In the article commonly recognised as the most representative of that period [11], it is assumed that every animal, independently of the species it belongs to, is capable of feeling pain: the mental state of pain is not species-specific. Therefore the identification of the mental state with a certain C-Fiber activation (or any other neural correlate) leads to the conclusion that all species should be found sharing the same neural structure and the same neural activation at the right moment. Even if we consider that parallel evolution might lead to the same physical structure, once the argument is extended to other psychological predicates (such as, for instance, hunger or sexual attraction), it becomes *overwhelmingly plausible* (Putnam's words) that these multiple realizations across species simply cannot be explained in terms of a theory grounded on the identity between mental and physical states. After all, even if parallel evolution could be proved in all known creatures, the conceivability of artificial silicon based systems capable of feeling pain, would definitively discard any attempt to establish an identity. Putnam's famous proposal is then to conceive a different approach to the mind, grounding it on the concept of a virtual machine analogous to the Turing Machine, but characterised by a few strategic differences.

It is useful to remind briefly what these devices are: a Turing machine (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. These processes may have the following schematic representation:

$$\{x_1, x_2, x_3, ...x_n\} \rightarrow A \rightarrow B \rightarrow C \rightarrow D \rightarrow ... \rightarrow [\textit{final status}]$$

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

This mechanism implies that 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, (e.g. starting from the functional state A the machine may change in favour of B with 30% of chances or C with 70%) then the machine is called Probabilistic Automaton. Finally, there are devices capable of processing sets of inputs in order to generate new sets of instructions: this ability allows simulating any possible TM generating a so-called Universal Turing Machine (UTM). In other words, the UTM is directly programmed by the input, which instructs the machine about the processes to apply thenceforth. The MRT assumes that the combination between a probabilistic automaton and a UTM gives in return a virtual device whose processes are consistent with the living beings' ones.

All these devices (TM, UTM and probabilistic automaton) are known as virtual machines because of their nature which makes them completely independent of any specific physical structure: it doesn't matter if the computation required by the set of instructions is performed by a neural system, a CPU or a series of cogs wheels. The focus is on the functional organization realized by the device (i.e. the instructions concerning its state transitions) and the functional state it can consequently reach, once the device has received a specific symbolic input. Furthermore, since the states are also independent, it is not even necessary for two systems to be functionally isomorphic (i.e. it is not necessary that they realize the same set of instructions) to reach the same state: different programs may lead to the same functional state.

In conclusion, the MRT entails that two generic neural structures A and B may realize a mental state M, but they can never be identified with the mental state itself: the relation between the physical system and its mental realizations is always contingent and there can be infinite physically different systems realizing the same mental state. The focus changes from the reductionist study of the neural correlate to the functionalist study of the realized functions[1].

Putnam's early argument has been originally applied to different neural structures belonging to different species, but few years later Jerry Fodor [7, 8] generalised the value of the MRT, presenting his assumption as the necessary consequence of Putnam's conclusions. The generalised version of the MRT has started appealing to the 70s studies on brain mapping and to the notions of neural degeneracy and plasticity: the key argument coming from these studies is that the nervous system of higher organisms is able to accomplish a single psychological task in a wide variety of ways by means of several neurological parts of the whole structure. As a consequence, the relation between physical and mental states proves to be contingent even when it is applied to the same species or a single neural system[2]: time becomes a legitimate variable to take into account when considering the contingency of the causal relation between the physical system (the implementer) and the functional state (the implemented).

2 The computability issue and the overestimation of the UTM

The superimposition of the processes performed by a virtual machine on the ones realized by cognitive organisms has been attractive since the very

[1]Subsequent articles (e.g. see [2] or i [12, §14]) have also dealt with the problem of the realization of more than a single functional state (or psychological predicate) at the same time. The solution proposed assumes complex living beings are able of realizing the processes of several virtual machines at the same time (i.e. in parallel).

[2]E.g. a single human being realizes the same mental state of pain during childhood and adulthood, despite the differences characterising the same neural structure in the two periods.

beginning: even those who have tried to discard the functionalist approach have rarely questioned the argument of the multiple realizations of mental states and have preferred to focus their attention on the implications the theory has on reductionism [5, 9, 10, 4]. A few exceptions are represented by those [17, 15, 1] who have challenged the likelihood of the argument by means of theoretical reasoning or stressing the failures of the predictions implied the generalised MRT. Nonetheless, I think a computational approach to this matter has been surprisingly ignored: the theory relies on the identification of the mind with the TM; should this identification be computationally inadequate, the MRT would be proved ill-grounded. As a matter of fact, there are three reasons that lead to this conclusion.

The first reason is the limited range of Turing-computable algorithms. To put it simple, the computational capacities of a TM are widely overestimated and they are usually erroneously attributed to Turing himself. There is a huge list of philosophical misconceptions about Turing's virtual machine [6] and they are all grounded on the erroneous assumption that in his articles Turing may have mathematically demonstrated how a UTM can compute any algorithm (i.e. the mathematical function that formally describes the set of instructions or program of the virtual machine) performed by any other machine with any architecture, given enough time and memory.

What Turing did demonstrate is that a UTM can realize any algorithm characterised by the following requirements (which define the 'mechanical method'):

1. finite number of exact instructions (each instruction expressed with a finite number of symbols) to make the machine change from one functional state to the following one.
2. Finite number of state transitions to produce the expected result.
3. In principle, a human being can carry it out only aided by paper and pencil.
4. It does not require insight or ingenuity to be carried out[3].

For the purpose of this article, it is sufficient to point out that the set of hypothetic algorithms realized by any TM is countable, that is to say, it is characterized by the same order of infinite of the integers. On the contrary, the number of all the hypothetic computable algorithms is uncountable (i.e. of a higher order of infinite): hence, there is an infinite number of algorithms

[3] These notions have a formal and rigourous equivalent[16, 3]: for the purpose of this paper it is sufficient to refer to their informal version.

which have a mathematical description and cannot be realized by a UTM, even if they are realized by differently structured systems.

If the algorithms implemented by neural systems are not found to meet at least one of the four requirements for Turing-computability, it must be concluded that a UTM may not simulate or even describe information processes in living beings. Consequently, it is necessary to study the way biological neural systems process their data, before formulating any hypothesis about the possibility to realize such processes by means of a virtual machine. Under these circumstances, the hypothesis of multiple realizations of mental processes may be empirically falsified: MRT cannot be established a priori.

It may be argued that even if we could find out that neural systems do not realize Turing-computable algorithms, 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 realizable algorithms, overcoming some of, if not all, the weak points of the classic machines.

Nonetheless, it seems that such a powerful virtual machine is unlikely to come and it is usually considered mathematically implausible[4]. Even if it were plausible, this objection would not lead far from the prospected path: these new hypothetic systems would not be asked to simulate a generic new set of algorithms but those specific of the parallel distributed -neural-systems. Once again, in order to be sure that the proper set of algorithms is part of the domain of these new machines (proving the soundness of MRT), it would be necessary to know beforehand what sort of algorithms are implemented by neural systems.

This conclusion leads to the second reasoning against the plausibility of the MRT. There is a particular causal relation between the physical structure of a neural system and the algorithm it implements: a neural network realizes a sheaf of sets of mathematical functions[5] defined by its architecture and by the computation performed by each single node of the network. The values assigned to the other variables, such as the weights of the synapses

[4]The existence and the features of devices that may result to be able to implement such Turing-incomputable algorithms have been debated at least for five decades. An essential bibliography and a brief account of this debate can be found in section two of the cited Copeland's article [6]. As a matter of fact, the probabilistic automaton already represents a virtual machine which is able to realize a wider set of algorithms, if compared to a TM. I mainly refer to the TM for the convenience of the reasoning, but the criticism is valid for the probabilistic automaton as well: the set of algorithms realized is still countable and the algorithms themselves are characterized by similar features.

[5]E.g. 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 the equations combined in single or multiple systems.

(i.e. the electrochemical conductibility of the synapses), fix the constants for any specific set of algorithms within this sheaf. Every modification in the architecture of the network or in the processes of the single nodes leads to a system that can or cannot solve a specific given task[6].

If we use simple connectionist models, the sheaf of algorithms implemented can be mathematically described with ease: in these conditions, the analysis of the relation between the neural structure and the implemented algorithm makes us conclude that the former has a causal influence on the latter. Nonetheless, 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 algorithms determined by the architecture, especially considering that, though extremely complex, single neurons compute their electrochemical signals in a way that can be described by adequate mathematical functions. In a few words, different neural systems realize different algorithms, require different amount of energy and time to perform the same task and -due to differences in vector conversion- differ in the way the information is encoded or stored, in the categories developed and in their resistance to physical damages. Thus, mathematical analysis of neural systems is telling us a different story from the one told by the MRT: in order to be able to process information -precisely- in the same way, two neural systems must be physically identical (i.e. two biological neural systems can hardly ever be functionally isomorphic due to the known structural differences across species and within the same one).

It is still possible to claim that whether or not two neural systems may perfectly match their processes implementing the same algorithm, this would not affect the hypothesis that a serial device may be conceived realizing neural processes. Once a probabilistic automaton were shown simulating the information processes of a neural system, the possibility to separate single states in the virtual machine would make it irrelevant for the MRT the whole second reasoning. Yet, the problem with this criticism is that it does not consider both the arguments so far described at the same time:

A. Whether or not a virtual machine may realize the set of instructions implemented by a neural system can only be established a posteriori.

B. The physical structure in neural systems is directly responsible for the processes implemented.

The two premises A and B lead inevitably to:

---

[6]The logical operator XOR is often cited in literature: it is known that there is no way to realize this computation with a single layer neural network (e.g. see [14, chap. 19, sect. 3].

C. In order to support an anti-reductionist path (MRT), it is necessary to use a reductionist strategy, seeking the knowledge concerning the processes realized by a neural system.

When everything is taken into consideration, the proof in favour of the multiple realizability of the mental states would be reached after it had become irrelevant.

The third reason against the plausibility of the MRT is grounded on the computational inadequacy of serial systems in simulating the unique features of biological neural systems. Biological systems deal with continuous and infinite inputs, processes and outputs, processing information in a flow; on the contrary, a virtual machine necessarily works with discrete and finite data and state transitions, following a step-by-step procedure. External data can reprogram a UTM to make it change its processes (once the input has changed the set of instructions, the device can also apply its rules to previously incomputable data), but the neural systems are able to change their processes both depending on and independently of the input. For instance, biological systems based on neural structures require a specific amount of energies in order to activate their systems: a lack of energy modifies the computational processes by means of a change in the computation performed in the single neurons of the network. This change takes place independently of both the awareness and the perception of such a lack in the organism This feature is not limited to the energy requirements: any physical alteration[7] directly modifies 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 can hardly give an account of these phenomena, despite the fact that they are very frequent in all living beings based on neural systems. Interestingly, Fodor [7] has used the argument of plasticity and degeneracy to propose his generalised version of the theory, but I think that this argument can be of use also against the virtual machine hypothesis, at least until these systems will be able to realize algorithms which can only be reprogrammed by input information.

Lastly, such differences make the parallel neural systems more robust in respect of time and energy 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 sufficient amount of time. On the contrary, the mechanical method implies that a serial system needs to follow all the given instructions

[7]E.g. structural damages or any other alteration of the neural architecture, chemical or electrical interference in electrochemical synapses, modification of the metabolic state of the neurons, etc.

in order to perform its transition among states: the lack of the time required to accomplish it would cause a failure in giving an output.

3 Making it through the MRT

It may be argued that it is here discussed the multiple realization of a whole set of instructions, but the object of the MRT is a single, independent and isolated functional state, which has its equivalent in the mental state/psychological predicate of a living being. Nonetheless, the supposed isolation of single psychological predicates such as pain, hunger, etc. is acceptable within the context of the known virtual machines, such as the UTM and the probabilistic automaton: these machines are characterised by serial processes and therefore allow the existence of autonomous functional states. Once the identification of the mind with virtual machines is disputed, the existence of states of this sort in the mind is challenged too: our self-beliefs about them may be misleading.

Let us push this line of thought a little farther. This article has outlined the following proportion:

Set of instruction: Turing machine = algorithm: system whose processes are mathematically describable

It may be argued that this proportion implies the following:

Functional state: Turing machine = assignation of values to all variables in the algorithm: system whose processes are mathematically describable

In the set of parallel neural systems (which is a subset of the mathematically describable systems), this proportion would imply that a particular kind of activation pattern would take the place of the third term in the second proportion. Though different from the 'C-nerve activation' correctly defined as *philosopher's fiction*[1], this would be anyway a completely theoretical object: a sort of photography of the entire structure, taking into account the whole network, the activation and metabolic status of all neurons and the disposition of every synapse to propagate its signals. Consequently, any change in any of the variables involved, would generate a different assignation to the variables as well as a different mental state, a conclusion that may seem to lead to an unusable theoretical object.

The problem is that biological neural networks are dynamical information processing systems, and consequently this perspective brings forth the concept of a theoretical object (the photography of the whole structure) characterised by an unavoidable incoherence. If the new definitions imply a concept of mental state which is both unusable and incoherent, then it seems it would be a good idea to discard the whole thesis, on the basis of its implications.

I think this is not a good reasoning: an analogy with the field of analysis in mathematics 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 general rules that can be applied to parallel neural systems (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), other rules that are species specific (the macro structure of the neural 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 [1], should make it possible to relate the new born theoretical mental states — indeed a dynamic concept, far from the static serial equivalent, but still usable- to the variances here described across species or within the single organism.

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 and the tool that allows such a comparison relies again in the mathematical description of the algorithms realised by the neural processes. Furthermore, there are many advantages in pursuing the use of this tool to understand mind processes. The algorithms describe the way every possible signal is computed by a 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 originally conceived by Putnam concerning the set of instructions of a probabilistic automaton, the specific study of the algorithms implemented by neural system would allow to describe every possible process these system perform in each of their layers, reaching important results in the understanding of the observable and hidden phenomena[8].

4 Conclusions

This paper states a methodological problem. There is no computational device able to realize all the uncountable possible algorithms: as a consequence, if the object of mind studies are the psychological predicates, it is necessary to study the specific processes that generate them. Whether or not these will result to be multiply realized, the computational study of

[8] Along this path, the main obstacle is represented by the epistemic indeterminacy due to the order of complexity of the biological neural systems, but I assume that grounding the models on the findings in neuroscience, a better explanatory value will be granted.

neural structures is the necessary first step of a realistic approach to the mind. Furthermore, contrary to what expected by the MRT, the more science gives us tools to investigate neural systems, the more it seems that the processes they implement are supervened by the physical matter and are characterised by a series of unique features.

Whenever 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. This procedure creates a useful analogy allowing an analysis narrowed to a part of the whole set of processes of the accessible system: as a consequence, the new aimed description is partial and indirect, because it refers to the supposed analogous system rather than to the original one.

My claim is that when multiple realizability is applied to neural systems, it is useful to conceive it as a tool giving access to incomplete descriptions of the psychological predicates: a similar constraint does not entail to discard the procedure 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 or if a better analogy is established (due to an accessible system which is closer to the unaccessible one), then the new description must be preferred to the partial one formerly achieved. In the field of mind studies, in the past few years, the mental processes are becoming more and more accessible and consequently new descriptions will be formalized thanks to this change: on this new ground, new explanatory theories will be built, showing substantial divergence if compared with the ones formerly inferred on the ground of the MRT.

In the attempt to save the MRT from Shapiro's remarks [15], Rosenberg has stated that this theory has been proposed to explain *the absence of discoverable psychophysical laws in a way compatible with physicalism*[13]. It seems today that we are moving towards the finding of these laws: should this happen by means of the mathematical description of the processes realised by the neural systems, the prediction here supported is that the multiple realizability tool will see the fields it has been applied so far restrained, in favour of the new tools.

BIBLIOGRAPHY

[1] W. Bechtel and J. Mundale. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 66:175–207, 1999.

[2] N. Block and J. Fodor. What psychological states are not. *Philosophical Review*, 81, 1972.

[3] A. Church. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58:345–363, 1936.

[4] P.M. Churchland. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78:67–90, 1981.
[5] P.M. Churchland. Functionalism at Forty: a critical retrospective. *Journal of Philosophy*, 1:33–50, 2005.
[6] J.B. Copeland. The Church-Turing thesis. *Stanford Encyclopaedia of Philosophy* [available online: http://plato.stanford.edu/entries/church-turing/], (last modified: August 2002, last consulted: July 2007).
[7] J. Fodor. Special sciences (or: on the disunity of science as a working hypothesis). *Synthese*, 28:97–115, 1974.
[8] Jerry Fodor. Special sciences: still autonomous after all these years. *Noûs*, 31:149–163, 1997.
[9] J. Kim. The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association*, 63:31–47, 1989.
[10] J. Kim. Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research*, 52:1–26, (1992).
[11] H. Putnam, Psychological predicates. In *Art, Mind and Religion*, pages 37–48. University of Pittsburgh Press, Pittsburgh 1967.
[12] H. Putnam. *Mind, Language and Reality. Philosophical Papers* vol. 2. Cambridge University Press, Cambridge, 1975.
[13] A. Rosenberg. On multiple realization and the special sciences. *The Journal of Philosophy*, 98:365–373, 2001.
[14] S. Russell and P, Norvig. *Artificial Intelligence: A Modern Approach*, Prentice-Hall, 1995.
[15] L. Shapiro. Multiple realization. *The Journal of Philosophy*, 97:635–654, 2000.
[16] A.M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, series 2, 42:230–265, 1936-37.
[17] N. Zangwill. Variable realization: Not proved. *The Philosophical Quarterly*, 42:214–219, 1992.