Consciousness, Functional Isomorphism and the Replacement Thought Experiment

Mauricio Eduardo Bieletto Bueno

University College London

Department of Philosophy

PhD Philosophy

I confirm that the work presented in this thesis is my own and the work of other persons is appropriately acknowledged.

Abstract

In this thesis, my main objective is the presentation and evaluation of several versions of a thought experiment that are offered for supporting a functionalist thesis concerning the generation of conscious phenomenal experiences. These versions depict replacement scenarios that describe an imaginary process for creating functional duplicates of the brains of conscious beings. It is assumed that the brain of a conscious person implements a determinate functional organization or structure, which can be initially understood as the abstract pattern of interaction between the different parts of the brain. The replacement process begins by adopting a certain level of functional organization by identifying a number of basic components in the brain that perform a certain function inside it. These basic *components* are then replaced by entities that may not share the same physiochemical composition of the original elements but that perform the same function inside the brain. When the replacement process is finished, the resulting system is a functional isomorph of the original system at the level initially determined. In this thesis, I will concentrate on two versions of the replacement thought experiment: the *neural* version and the *interchange* version. My first main claim is that the neural version of the replacement thought experiment, as presented in this thesis, gives adequate support to the thesis that the generation of conscious phenomenal experiences naturally supervenes on the property of instantiating the functional organization of the brain at a neural level. My second main claim is that a different version of the replacement thought experiment, which I call the interchange version, ultimately fails in supporting the thesis that the generation of conscious phenomenal experiences logically supervenes on the property of instantiating the functional organization of the brain.

Table of contents

Acknowledgments

I would like to thank my main supervisor, Lucy O'Brien, for offering me her valuable advice and helpful guidance. Also, I thank Paul Snowdon for discussing with me several of elements of this thesis. I want also to acknowledge the help provided by my uncle Jose and my aunt Graciela, and to my aunt Judy. Many thanks also to the CONACYT (Consejo Nacional de Ciencia y Tecnologia) for providing the scholarship that helped me to pursue my postgraduate studies at UCL.

Consciousness, Functional Isomorphism and the Replacement Thought Experiment

General Introduction

During the last thirty years, some philosophers have proposed a number of thought experiments that depict certain replacement scenarios describing a process in which a system that duplicates the functional organization of the brain is constructed. Among the authors that propose replacement strategies are Cuda (1985), Zuboff (1994), Kirk (1994), Chalmers (1996, 2010) and Tye (2006). Although the details of these thought experiments and their objectives varies with different versions of the thought experiment, in general the aim is to support a functionalist account of conscious phenomenal experiences. The replacement process described in these versions begins with the assumption that the brain of a conscious person implements a certain functional organization or structure, which can be initially understood as the abstract pattern of interaction between the different parts of the brain. In general, the process consists in determining a certain level of functional organization by identifying a number of basic elements in the brain that perform a certain function inside it. These basic elements are then replaced by entities that may not share the same physiochemical composition of the original elements but that perform the same function they have with respect to the original brain. In this way, when the replacement process is finished, the resulting system is a functional isomorph of the original system at the level initially determined.

There is a good justification for adopting the strategy of considering functional duplicates of the brain of conscious beings. After all, we have good empirical evidence in favour of the claim that the brain is the organ responsible for generating mentality. For instance, since the studies of the brains of aphasic patients made by Broca and Wernicke, we know that damage to some brain areas brings with it an impairment of certain mental functions.[1] However, and in spite of this empirical evidence, we do not know exactly what

1" In (2004), Rorden and Karnath evaluate the method of using brain injuries to infer mental function and discuss some of its limitations (including the questionable assumption that discrete brain modules deal with different mental functions). They argue, however, that complemented with new strategies for measuring brain

are the properties of the brain involved in the production of mentality. It is simply not evident how a system composed, among other things, by ions, aminoacids and water is capable of generating mental phenomena, like thoughts, memories or emotions. McGinn has famously expressed this worry as follows:

> We know that brains are the de facto causal basis of consciousness, but we have, it seems, no understanding whatever of how this can be so. It strikes us as miraculous, eerie, even faintly comic. Somehow, we feel, the water of the physical brain is turned into the wine of consciousness, but we draw a total blank on the nature of this conversion.[2]

The main topic of this thesis is precisely the problem of how a physical system can generate conscious phenomenal experiences. More precisely, I will examine a strategy that, according to several authors, supports a functionalist approach related to the generation of these experiences. According to this functionalist approach, mentality is generated or produced by the functional properties of the brain. Before I discuss these questions and the details of the replacement strategy, I want to briefly concentrate on the nature of the aspect of mentality known as phenomenal consciousness.

The proposal that the mental can be divided into two different aspects that are mutually autonomous or independent is well known and has been widely discussed. One of the main advocates of this distinction is Chalmers (1995, 1996, 2006). On the one hand, the mental involves all mental phenomena related to the production of behaviour. On the other hand, there is an aspect of the mental concerned with conscious subjective experience, with phenomenal consciousness, with the "what it is like" of experience.

> At the root of all this lie two quite distinct concepts of mind. The first is the *phenomenal* concept of mind. This is the concept of mind as conscious experience, and of a mental state as a consciously experienced mental state. This is the most perplexing aspect of mind and the aspect on which I will concentrate, but it does not exhaust the mental. The second is *the psychological* concept of mind. This

activity in healthy individuals, this method can still be very useful for scientific research. For an opposite point of view concerning the lesion method and the role of the brain in the generation of consciousness, see Majorek (2012).

2 " McGinn, (1989, p. 529)

> is the concept of mind as the causal or explanatory basis for behavior. A state is mental in this sense if it plays the right sort of causal role in the production of behavior, or at least plays an appropriate role in the explanation of behavior.[3]

This difference is conceived, first, as a metaphysical distinction. Mentality *per se* can be differentiated between the aspect involved to the production of behaviour, and the aspect involved to the phenomenal characteristics of an experience. This difference can also be conceived as an epistemological distinction. It also concerns the way in which the mental is understood and explained. Chalmers suggests that one of the effects of the division between the psychological and the phenomenal aspects of the mind is that the problems related to consciousness can be divided into an easy part and a hard part. Among the easy problems related to consciousness are how a physical system can react to external stimuli, how can it process information and generate inferences, how it can access its internal states and in general how it controls its own behaviour. In contrast, the hard problem of consciousness concerns how a system (in particular, the brain) can generate conscious phenomenal experiences. According to Nagel, although consciousness might be present at different levels or take several forms (for instance, there is surely a great difference between the conscious states that can be experienced by hominidae and the conscious states of rodents) conscious experience essentially involves a subjective element: that an organism is capable of having conscious experiences implies that there is something it is like to be that organism.

> ... the fact that an organism has conscious experience at all means, basically, that there is something it is like to be that organism [...] But fundamentally an organism has conscious mental states if and only if there is something that it is like to be that organism – something it is like for the organism.[4]

Several authors, including Nagel and Chalmers, suggest that the subjective character of conscious experience resists a physicalist or functionalist explanation. In particular, Chalmers (1995, 1996) has raised a number of arguments in favour of the claim that the

3 " Chalmers (1996, p. 11)

4 " Nagel (1974, p. 436)

phenomenal character of conscious experience cannot be reduced to physical or functional properties. One of the most discussed arguments against a functional approach of phenomenal consciousness concerns the logical possibility of a complete physical duplicate of a conscious being that lacks the capacity of having conscious phenomenal experiences. In contemporary philosophy, a being like this is known as a philosophical zombie.

In order to set up the problem I want to discuss in this thesis, I want to introduce first some reflections concerning the properties of physical models. A physical model is a device that is built in such a way that some of its properties mirror or represent the properties of another – perhaps much more complex – object. A very modest physical model of the Solar System made out of plastic balls and wires (like those created by school age children) may represent several interesting properties in spite of its simplicity. It could mirror, for instance, the respective size of the planets, the colour and shape of its surfaces and their order from the Sun (although such a model might not mirror all of these properties at the same time: a model of the Solar System representing the Sun as a sphere of 7cm together with a scale representation of the planetary orbits would be more than 30 meters long). More complex physical models of the Solar System include orreries and astrariums, which are mechanical devices that were used in the past for showing the motion of the planets around the Sun and, sometimes, even for predicting eclipses.

It is true, however, that these simple models cannot represent some other interesting properties of the bodies contained in the Solar System, like their gravitational interactions with other objects or the exact percentage of luminous radiation reflected by their surfaces. The motivations behind the construction of a physical model of determinate complexity depend on pragmatic grounds. Consider now a very simple physical model of the human brain of the kind that can be found in secondary schools. Depending of the level of complexity of the model – which can be made out from different materials, like plastic or wood – its parts can correspond, for instance, to the right and left hemispheres, to the visual cortex, the cerebellum, the hippocampus, etc. Like the simple models of the Solar System discussed in the preceding paragraph, the properties of these models represent only a very small subset of the properties of a real, biological brain. We cannot put these models under

a microscope and see miniature pieces of wood or plastic representing synaptic connections and neurons firing. The accuracy of these physical models depends, among other things, not only on our technical possibilities, but also on the particular objectives we have for building the model. It is not necessary to find a way to mirror the gravitational interaction that exists among all the bodies of the Solar System if the model is used for teaching school children. Analogously, it will be useful for a secondary school model of the brain to have properties that represent larger sections of the brain, but it surely does not need to mirror the dendritic structure of Purkinje neurons in the cerebellum.

Nevertheless, there does not seem to be any logical impediment to build a perfect physical duplicate of the brain of a conscious person. It is true that the technology needed for duplicating the causal properties of the brain – that is, the properties that allow it to generate mentality – is still unavailable, and it might be so for years to come. Nonetheless, building a perfect physical duplicate of the brain does not seem to imply any logical contradiction. It suffices to imagine a team of extremely competent neuroscientists possessing very advanced techniques for describing how the different elements of the brain work, together with a group of engineers that, with the help of the information provided by the neuroscientists, are capable of building physical devices able to duplicate the causal powers of the brain or parts of it.

A perfect physical duplicate of a conscious person would be indistinguishable from this person under any medical examination: they would have the same blood type, the same DNA, and will be prone to the same diseases. Note also that this duplicate will be *functionally* identical to the original subject: after receiving an external stimulus, it will process it in the same way as the original subject, and will produce the same behaviour. This duplicate will have internal physical processes that are functionally identical to the original being: for instance, the visual cortex of this physical duplicate will receive certain electrochemical signals from the lateral geniculate nucleus (which is the part of the brain that initially receives visual information from the retina) and in turn will send corresponding electrochemical signals to the areas in charge of the production of linguistic behaviour.

In order to illustrate the replacement process described by the replacement thought experiment, I would like to remember a widespread urban legend that gained a great amount of attention during 1969. According to this legend, Paul McCartney had died three years before in a car accident and was secretly replaced by a look-alike person. Some fans of the Beatles reported that hidden clues concerning Paul´s death could be found inside album cover imagery, the lyrics of some songs and even when one of the songs was played backwards. Fortunately, the legend about the death of Paul McCartney in a car crash and his posterior replacement by a look-alike person was false (as far as I know). However, it is not impossible to conceive a situation in which something strange happened to Paul in 1966 and he was secretly replaced not by a look alike person, but by something different.

Suppose that Paul effectively had a car accident in 1966, but in contrast with the original story, he survived the crash with apparently minor lesions in his head. However, when he was examined in the hospital, the physicians noticed that he started to develop a recently discovered but extremely harmful neural disease. The information possessed by the scientists concerning this illness was still minimal and inaccurate, but they at least knew that it was similar to motor neuron disease, a disorder that affects the control of voluntary muscle activity. This new disease, however, did not affect only the motor neurons, but all the other neurons in the brain of the patient. The prognosis, thus, was extremely unfavourable: the physicians told Paul that, unless he was subjected to a very complicated operation, he would fall in a comatose state in less than a month.

The operation suggested by the physicians is very simple to conceive (although some people doubted that the technology necessary for performing it was available at that time). Among the scientists there were expert neurologists that were able to determine, for each of the neurons inside Paul McCartney´s brain, their precise input-output behaviour. Experts in computer engineering used the information provided by the neuroscientists for constructing miniature computers capable of duplicating this behaviour. Take for instance the sensory neurons inside Paul´s brain, which are responsible for converting external stimuli from the environment (light, sound, temperature, etc.) into electrochemical signals that are in turn transmitted to other neurons inside the brain. The scientists provided these miniature

computers with tiny detectors for perceiving these external inputs, and programmed the computers in such a way that, if the biological neuron generates a certain electrochemical signal when received a determinate stimulus, the computer would generate exactly the same signal. Motor neurons and interneurons were built in the same fashion. Once that the physicians explained the process to Paul McCartney, and assured him that it was completely safe, he accepted to be operated.

Imagine now that the team of scientists isolate these miniature computers (we can call them from now “artificial neurons”) in the container of a fabulous operating machine, and put Paul McCartney inside it. The machine then removes, one by one, the biological neurons inside his brain, and replaces them with these artificial neurons. Once the artificial neuron takes its corresponding place, it will search for the neurons that were originally connected to the organic neuron and restore all the corresponding synaptic connections. The replacement process continues until all the neurons inside Paul´s brain have been replaced by artificial neurons. When the replacement process reaches this point, the result will be a being whose “brain” (let's now call it an “artificial brain”) is functionally identical to Paul´s organic brain at the neural level. From now, we will call this being “F-Paul”. Paul´s organic brain and F-Paul´s artificial brain are physical systems that have the same number of parts at the neural level, and these parts, in turn, process and generate the same electrochemical signals. I will not discuss, at this point of the thesis, how is that these artificial neurons duplicate the input output behaviour of the organic neurons inside Paul´s brain. For now, it will suffice to say that they will receive and transmit the same electrochemical signals that are interchanged between the original and the neighbouring neurons, and that is in this sense in which we say that artificial neurons are duplicates of the original neurons. Also, I will not discuss now the possibility of modifying the electrochemical nature of these signals. In the next chapter we will see the details of this modification and its consequences for the replacement process

This replacement process varies with respect to the different versions of the thought experiments: some versions consider the duplication of brain regions or structures

associated to particular mental abilities[5]. In other cases the process is described as involving the duplication of the input-output causal behaviour of the organic neurons of the brain by entities with a different physical makeup as these organic neurons, along with the preservation of their respective synaptic connections[6]. Other versions consider a mutual replacement or interchange of the physical realizers of the phenomenal conscious states of a person.[7] The common feature of all these thought experiments, however, is precisely the description of a replacement process in which the physicochemical properties of the original systems might be modified, while preserving, at the same time, their functional or structural properties.

It will be useful to compare this replacement operation with another one that, instead of replacing the damaged neurons inside Paul´s head by miniature computers, substitutes them with artificial systems that are – at the atomic level – identical to the organic neurons of Paul´s brain. Like the replacement process described earlier, there does not seem to be any logical impediment to build artificial organic neurons. It is true that the technology needed for doing it was not available during the sixties (when Paul had his accident), and most probably, it will not be available in the immediate future, if ever. Nonetheless, the point is that it is not logically inconceivable to imagine such operation. In this case, instead of a physical system functionally identical to Paul´s brain, the scientist constructed a physical system that – at the atomic level – is completely indistinguishable from Paul´s brain. Lets call this system P-Paul.

Perhaps it is easier to agree with the idea that P-Paul preserves the beliefs of Paul (for instance, that P-Paul believes that that the Beatles´ debut LP was *Please Please Me*, that the first drummer of the band was Pete Best and that he left the Beatles in 1962), that he is able to remember some of Paul McCartney's past experiences before 1966, or that he is able to feel cold, pain and the peculiar sensations associated with sneezing. After all, P-Paul was designed to be physically identical to Paul McCartney in 1966. At a biochemical level, they

5 " Zuboff, (1994)

6 " This is the strategy adopted in Cuda (1985), Kirk (1994) and Chalmers (1996)

7 " Tye (2006)

are completely identical. Moreover, P-Paul´s brain was originally created as being in exactly the same physical state as the brain of Paul McCartney when he died in 1966. So, P-Paul´s brain processes the information it receives from the external world in exactly the same way Paul McCartney's brain would have processed it.

In contrast, the doubts concerning F-Paul´s mentality rise mainly because instead of a biological brain, it has inside its skull a thing that – as scientists say – functions in exactly the same way as Paul McCartney´s brain. As the story was told, this explanation was initially very obscure and at that time very few people were able to make sense of it. What is for a non-biological physical entity – they asked – to function in exactly the same way as a biological human brain? After all, the scientist was very clear when he explained that the ersatz brain of F-Paul was not composed of biological neurons connected by synapses. He also hinted that, whatever the way the internal parts of F-Paul interchanged and processed information, there were no chemical neurotransmitters involved in this process. At a physicochemical level, Paul McCartney´s brain and the ersatz brain of F-Paul were two completely different things. But how two things whose internal physicochemical composition is so different "function" in exactly the same way? Notice that if it is shown that Paul´s functional duplicate, F-Paul, preserves the capacity of experiencing conscious phenomenal states, then his perfect physical duplicate, P-Paul, would preserve this capacity as well. The reason is that the functional properties of Paul´s brain would also be shared by P-Paul´s brain.

The aim of the replacement process just described was to create a functional duplicate of Paul McCartney´s brain. The point that interests me is whether the resulting system, F-Paul, is capable of having mental experiences. Does it share the memories of Paul McCartney? Is able to generate beliefs and to have desires? And more importantly, is it capable of experiencing conscious phenomenal states?

In the first chapter of this thesis I have three main objectives. The first objective will be to present and evaluate two famous thought experiments whose aim is to show that to duplicate the functional organization of the brain of a conscious person does not suffice for

generating mentality. These thought experiments are the Chinese Nation thought experiment (proposed originally in Block (2007b)) and the Brain Simulator thought experiment (proposed by Searle in (1980)). The second objective of the first chapter of the thesis is to formulate the functionalist theses supported by the replacement thought experiment. I claim that the best way for formulating these theses is with the notion of strong supervenience formulated in terms of quantifiers over possible worlds. This formulation will help to understand the differences among the theses that are supported by the several versions of the replacement thought experiment proposed by Cuda (1985), Kirk (1994), Zuboff (1994), Chalmers (1996, 2010) and Tye (2006). Finally, the third objective of the first chapter of the thesis will be the description of the variable elements of the replacement scenarios described in these versions.

In the second chapter of the thesis, I will present two initial versions of the replacement thought experiment. My objective is to determine whether these versions support at least the claim that the property of generating conscious phenomenal experiences *naturally* supervenes on the property of instantiating the functional organization of the brain at a determinate level. The first version based on the one proposed by Zuboff in (1994), contemplates a level of functional organization that corresponds to the regions of the brain associated to certain mental function, and that I will call the *regions-of-the-brain* version of the replacement thought experiment. I will then evaluate some objections concerning the properties of the entities that are proposed as replacements of these regions of the brain. The second version of the replacement thought experiment considers a more detailed level of functional organization of the brain: the neural level. This version is based on the thought experiments proposed by Cuda (1985), Kirk (1994) and Chalmers (1996). The replacement scenario described in this version contemplates the replacement of organic neurons by entities that duplicate the input-output relation that these neurons have with respect to the rest of the brain. The outcome of this replacement is a system that shares the functional organization of the brain precisely at the neural level. The argumentative strategy for claiming that the resulting system preserves the capacity of generating conscious phenomenal experiences will consist, first, in adopting the thesis of absent qualia as a sort of *reductio* hypothesis. The replacement process described in this version will generate a

series of cases in which, according to the *reductio* hypothesis, the conscious phenomenal properties of the system are eliminated. The strategy will be to argue that, in none of these cases, these conscious phenomenal experiences disappear. Since these are exhaustive cases, the conclusion will be that the absent qualia hypothesis is false. Finally, the last objective of this chapter will be to evaluate some further objections about the neural version of the replacement thought experiment.

In the third chapter of this thesis, I will present an objection presented originally by Searle in (1980) related to the idea that the implementation of a computational structure lacks objective conditions. This claim is known as the thesis of Universal Instantiation: for any computer program *C* and any sufficiently complex physical object *O*, there is a description of O under which it is implementing program *C*. This objection is relevant in the context of the replacement thought experiment for the following reason: in this thesis, I assume that the functional organization of a system – in particular, the functional organization of the brain at a given level – can be abstracted into computational abstract devices known as Combinatorial State Automata (CSA). This is originally proposed by Chalmers in (1995, 1996 and 1996b). The thesis of Universal Instantiation presents a challenge for those who support the claim that phenomenal consciousness supervenes on the functional organization implemented by a system. If this thesis of Universal Instantiation is true, almost any object can be described as implementing any computer program, and in particular, it can be described as implementing any CSA. Consequently, almost any object can be interpreted as implementing CSA_F, that is, the CSA that abstracts the functional organization of Paul´s brain at the neural level. Following Chalmers, I claim that, in spite of Searle´s objections, there are objective reasons for implementing a CSA.

In the fourth chapter of the thesis, I will discuss a different version of the replacement thought experiment whose objective is to show that the thesis of inverted qualia is, at least, empirically false. The version of the replacement thought experiment presented in the second chapter of this thesis might show that a system that preserves the functional organization of the brain at a neural level also preserves the capacity of generating conscious phenomenal experiences. However, the experiences generated by this

system might have a different qualitative character compared to the experiences generated by a biological brain. This possibility is illustrated by the inverted spectrum hypothesis, according to which the visual experiences generated by two isomorphic systems, P and Q, may differ in that the qualitative properties of the visual experiences generated by Q are phenomenally inverted with respect to the visual experiences generated by P. In this chapter I will also present an objection against this new version of the replacement thought experiment originally developed by Van Heuveln *et al* (1998) and Greenberg (1998). My claim in this chapter is that, in spite of these objections, this version of the replacement thought experiment successfully shows that the thesis of inverted qualia is, at least, naturally false.

Finally, in the fifth chapter of the thesis, I will discuss a further version of the replacement thought experiment originally proposed by Tye in (2006). Tye´s thought experiment envisages a mutual replacement, or interchange, of the internal states of two isomorphic beings. The aim of this thought experiment is to show that the thesis of absent qualia is *logically* false. More precisely, Tye´s strategy is to adopt a *reductio* hypothesis, according to which two systems A and B that are functional duplicates can differ in that A is capable of having conscious phenomenal experiences, while B is not. But if this is so, then the implementation of a certain functional organization is not sufficient for generating conscious phenomenal experiences. Tye´s strategy is to show that to assume the absent qualia hypothesis (understood in this sense) leads to a contradiction. Thus, if Tye´s arguments are sound, the absent qualia hypothesis is not just false: it is necessary false. Tye ´s argument, thus, can be seen as supporting a logical supervenience thesis: there is at least a functional organization F such that the property of generating conscious phenomenal experiences logically supervenes on the property of implementing functional organization F.

Chapter 1

The basic elements of the replacement scenario

Introduction

In this chapter, my first objective is the presentation and evaluation of two well-known thought experiments, presented originally by Block in (2007b) and Searle in (1980). The aim of these thought experiments is to show that the duplication of the functional properties of the brain is not a sufficient condition for creating a physical system that exhibits mental properties – intentionality in Searle's case, and consciousness in general in Block's. The first of these thought experiments is originally presented by Searle in the context of a reply to his famous Chinese Room argument: the Brain Simulator reply. Searle describes a system that duplicates the functional structure of the brain at a neural level. In this case, the duplication is performed by the implementation, by a physical system, of a computer program that simulates the way in which the brain of a native Chinese speaker functions when he understands stories written in Chinese and offers answers to questions about them. As we will see, Searle's strategy is to show that this program can be implemented by a system whose physicochemical composition seems to be obviously inadequate for generating intentional states, and in particular, for generating the conscious experience of understanding Chinese. The conclusion that Searle obtains from this thought experiment is that the simulation of the formal properties of the brain, at least at the neural level, is simply not sufficient for duplicating its causal properties, which are understood by Searle as the properties by which the brain generates intentionality, and in general, conscious phenomenal experiences. The second thought experiment – known as the “Chinese Nation” thought experiment – is developed by Block in (2007b). The scenario described by him is similar to the one presented by Searle. According to Block, what this thought experiment shows is that functionalism is a “liberal theory”, as it wrongly includes systems lacking mentality in the set of entities having it.

My second objective in this chapter is to explain and clarify the functionalist thesis that I claim is supported by a detailed version of the replacement thought experiment. This thesis can be initially (and informally) formulated by claiming that the instantiation of a certain functional organization by a physical system P is a sufficient condition for it to generate mental phenomena, and in particular, it is a sufficient condition for generating conscious phenomenal experiences. I want to suggest that the best way for expressing the functionalist thesis supported by the replacement thought experiment is to formulate it in terms of the notion of supervenience. More precisely, my aim is to show that it can be expressed by using the notion of strong supervenience formulated in terms of quantifiers over possible worlds. I briefly discuss the aims of some of the most known and discussed versions of the replacement argument, offered originally in Cuda (1985), Kirk (1994), Zuboff (1994), Chalmers (1996, 2006) and Tye (2006).

Finally, the third objective of this chapter is the description of the variable elements that compose the scenarios of the thought experiments on which these arguments are based. These elements are: (a) the modal character of the functionalist thesis supported by the different versions of the thought experiment; (b) the level of functional organization at play; (c) the nature of the replacement entities, and finally, (d) the way in which the replacement process is performed. As it has been mentioned in the general introduction of the thesis, a main feature of these scenarios is precisely the description of a substitution process that leads to the construction of functional duplicates of conscious beings. The details of this process, however, are not always described in the same way by the authors that have proposed arguments based on thought experiments of this kind. The modifications of the elements that constitute these scenarios give rise to several versions of the replacement argument. My aim in this section is to describe these elements in order to categorize the different versions of the replacement thought experiment that will be presented and evaluated in this thesis.

Section 1

Searle´s brain simulator and Block´s Chinese Nation

As it was explained in the introduction of this chapter, Block and Searle offer two well known thought experiments that depict a scenario in which the formal properties of the brain are duplicated in such a way that the resulting systems appear completely incapable of having any sort of conscious experiences. The objective of these thought experiments is precisely to show that this formal duplication is not sufficient for generating mentality. To be more precise, the aim is to claim that such duplication cannot generate consciousness, in Block's case, or intentionality in Searle's. This contrasts with what I claim is the general objective of the several versions of the replacement argument, namely, to show that the possession of certain functional properties – that is, the possession of certain functional organization – is a sufficient condition for mentality.

As we will see, Searle seems to justify his conclusion by appealing to the intuition that the systems envisaged in his thought experiment – giant networks of metallic pipes, for example – are too bizarre or unorthodox for having intentionality. It seems extremely doubtful that such systems, which in spite of their functional complexity have a very simple physicochemical composition, are able to match the causal properties of brains. Brains, in contrast, are extremely complex organic systems and the only objects in the universe that we have good reason to think are capable of generating intentionality. Meanwhile, Block does not rule out completely the suggestive force of this intuition, but also admits that it is not enough for justifying his claims, and suggests two further reasons for supporting the idea that a formal duplication of the brain at a neural level, by itself, cannot generate conscious experiences.

In order to get an adequate grasp of Searle's Brain Simulator, it will be useful to take a very brief look at the Chinese Room thought experiment. Remember that Searle's target, as he explicitly mentions it, is the thesis of Strong Artificial Intelligence, which can be understood as the claim that the implementation of a computational program is a sufficient

condition for having conscious understanding, and particularly, intentional states.[8] Inside the Chinese Room, a person manipulates symbols according to the rules contained in a program, rules that consider only the formal properties of these symbols, while excluding any semantic properties they may have. This manipulation is done in such a way that when external observers feed the room with questions written in Chinese, the person inside the room, after consulting the rules and performing the adequate symbolic manipulations, generates answers also written in Chinese that are in turn received by these external observers. Now, although the person inside the room may accurately follow all the rules specified by the program and perform all the required symbol manipulations, she is still clearly unable to understand a word of Chinese. But this is precisely what computational programs are: sets of rules that determine how to manipulate symbols according to their formal properties. From this, Searle argues that the implementation of a program is not a sufficient condition for having the conscious experience of understanding Chinese.

Now, contrary to some naive interpretations of the Chinese Room thought experiment, its objective is not to argue against the possibility of building artificial systems capable of generating mental phenomena, nor that machines are incapable of having intentionality. For Searle, the thesis that machines can think is true (although trivial), as well as the thesis that there are no logical impediments for building an artificial brain capable of generating mentality:

> … the issue about Strong AI is often taken to be the same as the question 'Could a machine think?' But of course the whole question is absurd. We are, after all, machines. If 'machine' is defined as any physical system capable of performing certain functions, then there is no question that humans and animal brains are machines. They are biological machines, but so what? There is no logical or philosophical reason why we could not duplicate the operation of a biological machine, using some artificial methods.[9]

So, the possibility of building an artificial machine capable of performing the function of a biological system is not something that can be derived from the Chinese Room thought

8 " Searle, (1980, p. 417)

9 " Searle (2002, p. 56)

experiment. How then are we to interpret the conclusions that Searle obtains from it? This can be explained if we consider Searle's Brain Simulator: note that the program described in the original Chinese Room thought experiment was designed for manipulating symbols already interpreted by external observers. But imagine now that somebody designed a computer program that simulates the neural structure of the brain of a Chinese speaker. The set of instructions that constitute the program mirror the way in which each neuron in the brain works, how they react when they receive an input signal, how they generate a corresponding output signal, and how they are connected to each other. The program can be implemented by an artificial physical system with the right characteristics (for instance, it would need to have a corresponding physical part implementing the input-output behaviour of each neuron in the original biological brain). External observers may feed the artificial system with stories and questions written in Chinese, and the system will produce, in turn, answers in perfect Chinese to these questions. This resulting system will behave thus exactly as the original Chinese Room. But there is of course a notorious internal difference. In this new case, the artificial system processes the Chinese stories in the same way as the brain of the Chinese speaker. While in the original Chinese Room the internal process was carried out by a person manipulating physical tokens of Chinese characters, in this new case the system is functionally analogous, at a very fine-grained level, to the brain of a genuine Chines speaker. The question raised by this situation, thus, is whether the Brain Simulator understands Chinese in the same way as the original speaker.

Not unexpectedly, Searle's answer is negative. He argues as follows: suppose that the person inside the Chinese Room, instead of processing Chinese characters according to their formal properties, is implementing a computer program that allows him to control the valves of a very complex system of water pipes. Valves connect these pipes, and for each of the synapses inside the brain of the Chinese speaker there is a corresponding valve, in such a way that this system of water pipes is a precise simulation of the brain of the Chinese speaker at the neural level. When this complex system receives certain information as an input – which may be constituted, perhaps, by a series of Chinese characters forming a question about a story – there is a transducer whose task is transform this information into water signals. The system processes these signals according to the rules of the program,

rules that mirror the way that the brain of the Chinese speaker processes this same information. It then produces an output that, after being translated by the transducer, can be interpreted as an answer to a question about the Chinese story. Nonetheless, the implementation of the program just described does not allow the man to understand the Chinese story, and it would be equally mistaken – Searle argues – to claim that the water pipes, or the combination of the man and the water pipes, are able to understand the story.

There are two points concerning Searle´s thought experiment that need to be clarified. First, the Chinese Room thought experiment, as Searle originally presents it, is conceived as a challenge against the claim that intentionality can be generated by the implementation of a computer program. But at least in its original version, Searle´s thought experiment does not seem to be directly related to the claim that computation suffices for consciousness. However, notice that it will be easy to build a similar thought experiment that involves consciousness instead of intentionality. This thought experiment might describe a program such that its implementation by the man inside the room allows it to be conscious of its environment, or even to experience conscious phenomenal states. Moreover, Searle argues that there is a conceptual connection between consciousness and intentionality. This connection implies that, in order to build a theory of intentionality, it is necessary first to give an account concerning the nature of consciousness. Consciousness, thus, is at the core of Searle´s perspective.

> Only a being that could have conscious internal states could have intentional states after all, and every unconscious intentional state is at least potentially conscious. This thesis has enormous consequences for the study of the mind. It implies, for example, that any discussion of intentionality that leaves out the question of consciousness will be incomplete.[10]

Second, it might not be entirely clear, at this moment, the relation between implementing a computer program and having a certain functional organization. However, this relation will be clarified later in this chapter, where I discuss the notion of a

10 " Searle (1992, p. 132)

Combinatorial State Automaton (proposed originally by Chalmers in (1996 and 1996b)) and the conditions that a physical system needs to satisfy in order to implement it.

Before evaluating Searle's argument, it will be useful to consider a very similar thought experiment offered by Block in (2007b). Block's aim is to show that functionalism is guilty of what he calls “liberalism”: systems lacking mental properties are wrongly classified by functionalism as having them. Imagine a physical entity that externally looks like a human being capable of having mental experiences like any other normal human being. Suppose that from the point of view of an external observer, this entity seems to be identical to Paul McCartney: it behaves like Paul, it has his voice and is a very good bass player. Nonetheless, from an internal perspective, this entity and Paul McCartney are quite different. To start with, it does not have a brain, or at least, it does not have anything that can be classified as a biological brain. Instead, there is a cavity inside its head in which there are a number of tiny homunculi, one for each of the neurons inside Paul McCartney's brain. To each neuron will correspond a machine table composed by a set of quadruples describing the current state of the neuron, the input received, the next state of the neuron and the corresponding output. What these homunculi do is to implement each of these machine tables: they receive an input signal, and by following the instructions of the machine table, they change the state of the neuron and produce a corresponding output. In such a way, this complex of homunculi manages to implement the functional organization of Paul McCartney at the neural level.

Although it does not seem obvious that the entity just described lacks mentality, Block also suggests a nomologically possible version of this thought experiment in which the absence of mental properties seems to be much more evident. This thought experiment is known in the philosophical literature as the “Chinese Nation”. Suppose that all Chinese citizens are equipped with a radio for interchanging linguistic information with other Chinese people and with an artificial body kept in a laboratory in Beijing. This network of people plays the role of an external brain connected with this body. Each person plays the role of one neuron of this artificial brain by interchanging verbal signals, in a similar way as the homunculi already described implement a determinate task. Block claims that

although this system can implement the functional organization of an intelligent individual, it does not have mentality at all. Having certain functional structure, therefore, does not suffice for having mentality. If the objection is successful, then functionalism is a liberalist theory, and does not give an adequate answer to the question about the nature of mental states.

> What makes the homunculi-headed system […] just described a prima facie counterexample to (machine) functionalism is that there is prima facie doubt whether it has any mental states at all – especially whether it has what philosophers have variously called "qualitative states", "raw feels" or "immediate phenomenological qualities."[11]

What does justify the claim that the systems envisaged by Searle and Block are not able to have intentionality or experience conscious mental states? Surely, there seems to be an intuitive resistance concerning the idea that a bunch of wet metallic pipes, in spite of the complexity of the system they form a part, may have the same capacities as the brain of a human being able to understand stories written in Chinese, English or any other language, as well as having other sort of conscious states, like experiencing an emotion or feeling pain. Equally, a huge crowd of people using radios and speaking among themselves does not look to be the right candidate for having genuine mental states, regardless of the fact that it has the same functional properties as the brain of an intelligent being. There are several physical systems that have an enormous number of parts and exhibit a very complex functional organization. Electronic computers, highly complicated watches, or even cities, planets or galaxies, have a large number of parts that interact in very intricate ways, but as complex as these systems might be, we simply do not attribute them the ability of having conscious experiences.

But even if these intuitions are widespread, it is simply not obvious why a system made of biological components and not a bunch of metallic pipes and water can support mentality. After all, at a microscopic level there is only an extremely complex system of brain cells interchanging electrochemical signals. Although it is true that the only systems

11 " Block (2007b, p. 43)

we know can support mentality are precisely biological brains, to affirm that systems with similar complexity cannot have mentality just because their physical composition seems to be very simple – rusty metallic pipes and water – does not appear entirely convincing. In fact, Searle himself leaves open the question whether to build a conscious artificial system, made out from non-organic components, is a real possibility [12], although the current empirical knowledge we have does not allow us to give a decisive answer to this question yet. Searle's main point, however, is that simulating just the formal structure of the brain is not sufficient for creating an artificial thinker.

> The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states.[13]

So according to Searle, this simulation falls short of mimicking what it is really important about the brain: its causal properties. But what are these causal properties? As Searle explains them, they concern the power of the brain to produce intentional states, and a functional simulation of the brain cannot reach this power or capacity. But again, there does not seem to be a reason why this simulation cannot have these causal powers, apart from the weak intuition that the physical composition of the system implementing this simulation is different from the brain.

Block's objections against functionalism seem to rely on the intuition that these bizarre simulations of the brain at a neural level are simply not the right kind of entities to which mentality can be attributed. Nonetheless, he realizes that without further arguments, this intuition is far from being decisive, so he offers two more reasons to reinforce it. First, although it is true that there is not a definitive theory that explains how is it that beings with

12 Cf. Searle, (2002, p. 56): "There is no question that an artificially made machine could, in principle, think. Just as we can build an artificial heart, so there is not reason why we could not build an artificial brain. The point, however, is that any such artificial machine would have to duplicate, and not merely simulate, the causal powers of the original biological machine".

13" Searle (1980, p. 421)

brains are able to experience conscious qualitative states, we are sure that brain-headed entities – and in particular, human beings – are able to experience these states. Thus, it seems to be a mistake to affirm that, since it is not clear how the systems described by Block and Searle may support mentality and the ability of experience qualitative states, it is equally unclear how is that brains are able to do the same. There is good empirical evidence that brain-headed beings are capable of being conscious. It might be possible to reinforce this idea by noticing that the only systems we currently know are capable of generating consciousness are, precisely, biological brains. Until now, we have not found systems able to support mentality in environments hostile to organic beings. As far as we know, there are no such systems in the Moon, or in the Martian deserts, or even in some terrestrial regions. But if mentality depended on functional structure and not on the physicochemical or biological properties, why have we not found beings with mentality in these environments?

Second, both the Brain Simulator and the Chinese Nation described respectively by Searle and Block are designed to mimic beings with mental properties. In contrast, human beings and their brains are not designed to mimic anything. This fact might show that it is not necessary to explain the behaviour of these simulations by appealing to mental entities, like beliefs or desires. As Block notes "The best explanation of the homunculi-heads' screams and winces is not their pains, but that they were designed to mimic our screams and winces".[14]

According to Block, his objection against functionalism depends on the following claim: there is a theory that seems to imply an absurd conclusion (i.e., that the China-Body system is capable of generating conscious states). If it cannot be shown that the theory does not imply this conclusion, or if there is no way of explaining why the conclusion is not as absurd as it seems to be, then we have reasons for *not* accepting the theory.

> I claim that there is no independent reason to believe in the mentality of the homunculi-head, and I know of no way of explaining away the absurdity of the conclusion that it has mentality (though of course, my argument is vulnerable to the introduction of such an explanation).[15]

14 " Block, (2007b, p. 77)

15 " Block (2007b, p. 77)

I suggest that the replacement thought experiment can be interpreted as offering the independent reason requested by Block. But how can then be explained away the perceived absurdity of claiming that Searle´s brain simulator or Block´s China-Body system are capable of generating conscious phenomenal states? Tye (2006) compares this case with a mathematical conjecture that has not been proven. Perhaps most mathematicians suspect that the conjecture is false, and in absence of a proof, there does not seem to be any contradiction in imagining its falsity. But suppose that the conjecture in fact has a proof not discovered by any mathematician. In this case, the conjecture would be true in spite of these suspicions. Analogously, Tye suggest that " it may well seem to me that I can imagine a homunculus-headed system that duplicates a normal human functionally and yet lacks qualia. But in reality this is not conceptually possible."[16]

Section 2
The notion of supervenience

Remember again the initial version of the thought experiment presented in the introduction of the thesis where an imaginary procedure for creating a functional twin B of a physical system A capable of generating conscious phenomenal properties is described. The procedure consists in identifying a number of basic elements that belong to system A and replacing them by other entities that do not share the same physiochemical composition, but that preserve the input-output behaviour of these basic elements. Now, if it is successfully argued that the property of system A of generating phenomenal is preserved along all the replacement steps of the process – including, of course, the last step, which is precisely system B – then what has been shown is that the physical properties of system A are irrelevant for the generation of the property of generating conscious phenomenal experiences.

This is an important result. However, note that there might be certain non-functional properties that were preserved along the replacement process, and there is simply no way

16 " Tye, (2006, p. 161)

for saying that they are not involved in the generation of phenomenal consciousness. What has to be done in order to support the claim that conscious phenomenal properties supervene on functional properties is to assure that the only properties shared by systems A and B are functional properties. More precisely, we need to be sure that the only property shared by systems A and B is the property of instantiating the same functional organization. I will argue later that there is a naive way of conceiving the replacement thought experiment that inadvertently preserves other non-functional properties of system *A*. I am not claiming that the authors of published versions of the thought experiment commit this mistake. However, it is easy to interpret the thought experiment in this way. For the moment, imagine that the replacement process preserved just the property of instantiating the same functional organization. In this case, if it is shown that the capacity of system A of generating conscious phenomenal experiences are preserved along the entire process, then we have shown that the implementation of the functional organization of system A at the level initially established is a sufficient condition for the generation of these experiences. Call this thesis the sufficiency thesis:

> **[Sufficiency thesis]:** there is a set *F* of functional organizations such that the implementation of a member of set *F* by a physical system is a sufficient condition for it to generate conscious phenomenal experiences.

My claim in this section is that the sufficiency thesis is better expressed in terms of the notion of supervenience. To appeal to the notion of supervenience for giving an account to the general aims of replacement arguments has several advantages as this notion can be understood as a formalization of the claim that one set of facts depends on, or is determined, by another set of facts. Take into consideration some initial examples: we know that the property of a circle of having an area of such and such dimensions supervenes on the length of its radius. In other words, it is not possible for two circles to differ with respect to the size of their areas without differing with respect to the length of their radii. Also, the property of a physical body of having a certain weight over the surface of the Earth supervenes on its mass. No two physical bodies can have a different weight over the surface of the Earth if they have the same mass. Similarly, the property of belonging to the

crew of the Apollo 13 supervenes on the property of belonging to the set {James Lovell, Ken Mattingly, Fred Haise}. Belonging to this crew depends on being a member of the latter set, so it is not possible for two persons to be different with respect to the property of belonging to the crew of the Apollo 13 without being different with respect to the property of belonging to this set.

Consider the strategy behind the replacement thought experiment: we imagine a process in which the basic elements of a physical system *A* capable of generating mentality are replaced by other entities in such a way that the functional organization of system *A* is preserved along the process. The outcome of this process is a system *B* whose physicochemical properties may differ from the ones of system *A*, but that shares the same functional organization. In the example described in the introduction, the system at play is Paul McCartney´s brain. In this case, the basic elements are its organic neurons, which were replaced by artificial neurons capable of duplicating the input-output behaviour of these organic neurons. The result is the creation of a non-biological system that only shares with Paul McCartney's original brain the property of instantiating the same functional organization. The crucial step of the argumentative strategy is to show that the conscious phenomenal properties of Paul are preserved along the replacement process. But if this true, what has been shown is the following conditional sentence: if there is no difference concerning the functional properties of these systems, then there is no difference concerning their conscious phenomenal properties. This in turn means that conscious phenomenal properties supervene on functional properties. The relation of supervenience can be better expressed as follows:

> (S) There is a set *F* of functional organizations such that the property of a system P of generating conscious phenomenal experiences supervenes on P instantiating a member of set *F*.

One of the problems with suggesting that the conclusion of the replacement argument can be expressed with the help of the notion of supervenience is that there are, at least, two

non-equivalent notions of supervenience at play: weak supervenience and strong supervenience, where the latter logically implies the former, but not conversely. Also, both weak and strong supervenience can be expressed by using modal operators and quantifiers over possible worlds. In (1993 and 1990) Kim identified these two formulations of the notion of supervenience and suggested their mutual equivalence, although McLaughlin (1995) showed that both weak and strong supervenience formulated in terms of modal operators are stronger than the formulations in terms of quantifiers over possible worlds. Lets start with the notion of weak supervenience as formulated by modal operators. We say that a set *M* of mental properties supervenes on a set *F* of functional properties when necessarily, if anything has a mental property *P* that belongs to set *M*, then there is a functional property that belongs to set *F* that it also has, and in fact, if anything has that functional property, then it has that mental property. Formally, the modal operator version of weak supervenience can be expressed as follows:

$$(\mathrm{WS_M})\ \Box\ \forall x\ \forall M' \in M\ [M\exists x \rightarrow \exists F\ \exists\in F\ (F\ \exists x\ \&\ \forall y\ (F\ \exists y \rightarrow M\ \exists y))]$$

Weak supervenience can also be formulated in terms of quantifiers over possible words:

$$(\mathrm{WS_Q})\ \forall w \forall x \forall y((x \text{ is in } w \text{ and } y \text{ is in } w) \rightarrow (\forall F'(F'x \leftrightarrow F'y) \rightarrow \forall M'(M'x \leftrightarrow M'y)))$$

According to the quantifier version of weak supervenience, we say that *M*-properties weakly supervene on *F*-properties if for any possible world *w* and any individuals *x* and *y* in *w*, if *x* and *y* are *F*-indiscernible in *w*, then they are *M*-indiscernible in *w*. As Kim realizes, to say that a set *A* of properties weakly supervenes on a set *B* of properties is not sufficient to say that *B*-properties determine or fix the *A*-properties[17]. The only requirement for weak supervenience is that there are no two objects with the same *B* properties that differ in their *A*-properties within the same possible world. However, weak supervenience is compatible with the existence of an object in another possible world that has the same *B*

17 " Kim (1993, p. 60)

properties of an object in this world, but without the same *A* properties. This is not allowed by strong supervenience, which is formally expressed as follows:

(SS_M) $\Box \forall x \, \forall P \in M \, [Px \rightarrow \exists Q \in B \, (Qx \, \& \, \Box \, \forall y \, (Qy \rightarrow Py))]$

Strong supervenience can also be formulated in terms of quantifiers over possible worlds:

$(SS_Q)\forall w_1 \forall w_2 \forall x \forall y((x \text{ is in } w_1 \text{ and } y \text{ is in } w_2) \rightarrow (\forall F' \, (F' x \leftrightarrow F' y) \rightarrow \forall M' \, (M' x \leftrightarrow M' y)))$

The difference between the modal operators and the quantifier formulation of weak and strong supervenience will be clearer later. Meanwhile, it is important to notice that most, if not all versions of the replacement thought experiment are offered for giving support to the claim that the functional properties of a system fix or determine its mental properties. It is for this reason that the most adequate notion of supervenience for our purposes is, precisely, strong supervenience. This thesis can be expressed as follows:

> (SS) There is a set *F* of functional organizations such that the property of a system P of generating conscious phenomenal experiences *strongly* supervenes on P instantiating a member of set *F*.

Set *F* might be a set that includes only one element, for instance, the functional organization of the brain at a neural level. But we will see that other thought experiments that depict a replacement scenario consider different levels of functional organization. It might be argued, for instance, that the human brain instantiate a certain functional organization at the *regions-of-the-brain-level*. In the next chapter, I will discuss a version of the replacement thought experiment that considers this functional level.

There is a problem with the formulation of supervenience in terms of modal operators, which makes it unsuitable for expressing the general form of the conclusion of the replacement argument. According to the original sufficiency thesis, if a physical system P implements functional organization F, then it will be able to experience conscious

phenomenal states. Note that the direction of the conditional is important: the Sufficiency Thesis is consistent with the existence of conscious beings that may lack any sort of functional organization – for instance, immaterial beings that may not have any identifiable physical parts – but that exhibit mental properties. As an example, we may suppose that this is what happens to entities like the ghost of the King of Denmark or the archangel Gabriel. If such entities exist, perhaps their mental properties are generated in virtue of unknown processes not related to the functional properties they might posses (perhaps there is no way for determining what these functional properties are, if any) or in the archangel case, perhaps by some sort of miracle or by divine intervention. Nonetheless, understood in this sense, ghosts and archangels are clearly not incompatible with the Sufficiency Thesis. We are concerned only with the question whether functional twins – systems that share the same functional organization – are also mental twins. For these reasons, both formulations of weak supervenience are inadequate for expressing the general conclusion of the replacement argument: they imply that it is necessary that if something has a mental property, then it has a functional property. Thus, if the formulation of strong supervenience in terms of modal operators is adopted as the adequate formulation of the conclusion of the replacement argument, we run the risk of unjustifiably ruling out the existence of beings like the ones described earlier. Of course, my intention here is not to say that there are good reasons for saying that ghosts or archangels really exist. I am just simply saying that the replacement argument should be seen as neutral with respect to this issue, and for this reason, to formulate its general aim by using a notion which rules these entities out is inadequate.

Kirk (1994) mentions an analogous problem involving the formulation of minimal physicalism. He argues that minimal physicalism commits us to what he calls the Strict Implication Thesis, which can be formulated as follows: assume first that the universe is physically closed, that is, the totality of physical events are determined by physical laws. Let P be the set of all true statements formulated in an idealized physics. These statements describe all physical situations at every place and time. Assume now that Q is another set of true statements that ascribe mental states to beings in the universe that is specified by the statements of set P. With this in mind, we can express the Strict Implication Thesis: P

strictly implies Q: it is impossible that all the members of P should be true and some members of Q false. The intuitive idea behind the Strict Implication Thesis is that mental facts are determined by physical facts. Now, Kirk suggests that the Strict Implication Thesis is more suitable for expressing minimal physicalism than strong supervenience because strong supervenience does not allow the possibility of Cartesian worlds in which there are non-physical minds.

> [supervenience] is unsuitable for minimal physicalism because it entails that necessarily (that is, in any possible world) any given mental property is correlated with some physical property. The idea is that there can be no mental difference without a physical difference. And that says too much for minimal physicalism because it rules out worlds where certain kinds of Cartesian dualism reigns. […] In contrast, the Strict Implication thesis leaves it open whether there might, logically, be such Cartesian worlds.[18]

And also:

> … strong supervenience may perhaps be an appropriate relation if you intend to use it to ascribe a purely physical 'nature' to the mental. But the Strict Implication thesis doesn't rule out the possibility that some minds should have been non-physical. [19]

Kirk's worries are similar to the ones I mentioned earlier. In this case, if the functionalist thesis is expressed by using the notion of strong supervenience formulated in terms of modal operators, we are at risk of not allowing these Cartesian worlds that may contain entities lacking functional properties that are able to experience conscious states, for example. However, I think that – at least with respect to the formulation of the general aims of the replacement argument – it is possible to use an alternative notion of strong supervenience formulated not in terms of modal operators, but in terms of possible worlds.

18" Kirk (1994, p. 81)

19" Kirk (1994, pp. 81-82)

(SS_Q): $\forall w1\ \forall w2\ \forall x\ \forall y$ ((x is in w_1 and y is in w_2) → ($\forall F'$ ($F'x \leftrightarrow F'y$) → $\forall M'(M'x \leftrightarrow M'y)$))

What this formulation says is that, for any possible worlds w_1 and w_2, and for any objects x and y, if x has in world w_1 the same functional properties that y has in w_2, then x has in w_1 the same mental properties that y has in w_2. What this formulation says is that if two objects are P-indiscernible, that is, if they have the same functional properties, then they are M-indiscernible, that is, they have the same mental properties. But note that accepting it does not commit us to accept that the possession of a mental property implies the possession of a physical (or functional) property. Thus, this formulation does not rule out the possibility that some minds are not physical, like the archangel Gabriel or the ghost of the King of Denmark.

Section 3

The variable elements of the replacement scenario

In this section, I will enumerate some of the variable elements that define the scenarios of some of the most discussed versions of this thought experiment. The variable elements of the replacement scenario I will consider in this section are the following:

(a) The modal character of the functionalist thesis supported by the different versions of the thought experiment
(b) The level of functional organization at play
(c) The features of the replacement entities
(d) The features of the replaced entities

(a) The modal character of the functionalist thesis

One difficulty of using the notion of supervenience concerns the various versions of the replacement thought experiment that have been offered in the literature. As they are explicitly formulated, the theses that the authors claim are supported by their respective

versions of the thought experiment show important variations, and in some cases these authors have very different positions with respect to the nature of mental states. These differences may motivate the view that there is not a unique way for expressing the functional thesis supported by the various versions of the thought experiment, and that the best strategy is simply to identify some similarities without trying to identify a single formulation of it.

Thesis (SS) can be understood in two different ways, which corresponds to a further distinction between natural and logical supervenience. As the claim is that the formulation of the thesis supported by the replacement thought experiment can be expressed with the notion of strong supervenience formulated in terms of quantifiers over possible worlds, I will use this formulation here. Both natural and logical supervenience can be formally expressed as before:

(SS_Q) $\forall w1\ \forall w2\ \forall x\ \forall y$ ((x is in w_1 and y is in w_2) $\rightarrow$ ($\forall F(Fx \leftrightarrow Fy) \rightarrow \forall M(Mx \leftrightarrow My)$))

The difference consists in that the universe of discourse of the first two quantifiers ranges on all possible worlds, in the case of logical supervenience, and on only the natural possible worlds, in the case of natural supervenience. Lets consider first natural supervenience. To say that the property of generating conscious mental states supervenes naturally on the instantiation of a certain functional property – that is, on the property of implementing the functional organization of the brain at a neural level – means that in no naturally possible world are there two systems with the same functional organization but differ in that only one of them has the property of generating conscious mental states, while the other does not:

> (Natural Supervenience) The property of generating conscious mental states naturally supervenes on the property of implementing functional organization F.

In this reading, the universe of discourse of the first two universal quantifiers ranges on all naturally possible worlds.

(Logical Supervenience) The property of generating conscious mental states logically supervenes on the property of implementing functional organization F.

In this reading, the universe of discourse of the first two universal quantifiers ranges on all possible worlds, natural or not. More precisely, the theses of natural and logical supervenience will be expressed as follows:

(NSS) There is a set F of functional organizations such that the property of a system P of generating conscious phenomenal experiences *naturally* strongly supervenes on P instantiating a member of set F

(LSS) There is a set F of functional organizations such that the property of a system P of generating conscious phenomenal experiences *logically* strongly supervenes on P instantiating a member of set F.

The difference between natural and logical supervenience will help us to categorize the several versions of the replacement thought experiment. The first category corresponds to those versions of the replacement thought experiment that support thesis (NSS). A clear example is the version offered by Chalmers in (1996, 2006) This version is explicitly formulated as supporting the Principle of Organizational Invariance:

(Principle of Organizational Invariance): Given any system that has conscious experiences, then a system that has the same functional organization will also have conscious phenomenal states.

Chalmers offers two different versions of the replacement argument. Their main objective is to argue against two similar hypotheses concerning the way in which qualitative conscious states are generated. The first is the absent qualia hypothesis. According to it, it is possible for two physical systems A and B to share the same functional organization at a neural level, while at the same time differing in that A is able to

experience conscious phenomenal states, but B is not. Arguments that support the Principle of Organizational Invariance are offered against the claim that isomorphic systems do not share the same mental properties. Nonetheless, it has been suggested that systems that have the same functional organization may not share the *same* conscious experiences. For instance, two beings that are functionally equivalent may differ in that the visual experiences of one of them are inverted with respect to the visual experiences of the other. One member of the pair would have yellow visual experiences when he is looking at a blue sky, or green visual experiences when seeing a strawberry, while the other would have the visual experiences that these objects normally produce. But of this is so, then having the same functional organization as a conscious being may suffice for being conscious, but not for having the same conscious states. This is the Inverted Qualia hypothesis: it is possible for two physical systems A and B to share the same functional organization at a neural level but whose conscious phenomenal experiences – in particular, their qualitative experiences – are inverted with respect to each other.

Chalmers has a complex position: while he clearly disagrees with the idea that phenomenal consciousness supervenes on the physical, he distinguishes between natural and logical supervenience. He explicitly rejects the claim that phenomenal consciousness supervenes logically on the physical: according to him, it is perfectly conceivable a system that is a complete physical duplicate of a human being, but lacking phenomenal consciousness (or in other words, he accepts that zombies are possible). Nonetheless, he also accepts the idea that phenomenal consciousness supervenes nomologically on the physical.

The second category corresponds to the versions of the replacement thought experiment that can be interpreted as supporting thesis (LSS). Among these are the versions proposed by Zuboff (1994) and Tye (2006). Zuboff´s arguments can be interpreted as arguing against both the absent qualia hypothesis and the inverted qualia hypothesis. With regard to the absent qualia hypothesis, Zuboff argues that it is possible to know a priori that the replacement of a region of the brain that preserves its original functional role will also

preserve the particular nature of the experience formerly controlled by that region of the brain.

> So we can know a *priori* that the preservation of nothing more than that brain chunk's extrinsic causal role within the rest of the mental system also perfectly preserved all the nature of any experience to which that chunk of brain had made a contribution.[20]

With regard to the inverted qualia hypothesis, he claims that we can know a priori that the replacement of a region of the brain that preserves its functional role also preserves the same experiences formerly produced by that region of the brain.

> ... honestly speaking about and behaving towards colors as though they looked the same on both sides of the visual field based on an experience of them as radically different is an impossibility, a contradiction [...] The sameness of function must logically determine the sameness of experience.[21]

According to Zuboff, the claim that the replacement of the visual cortex of a conscious being – like Paul McCartney – by a gadget that preserves its same functional properties does not preserve the same visual experiences is not only false, but also contradictory. Tye´s position is similar to Zuboff´s. The thought experiment proposed by Tye in (2006) and the arguments he develops from it are explicitly directed to give an answer to the absent qualia hypothesis. Tye formulates this hypothesis as posing an objection to functionalism, and more precisely, as an objection to a functionalist account of phenomenal consciousness. Tye´s strategy consists in showing that the absent qualia hypothesis – understood as the claim that functional twins can differ in that only one of them has conscious phenomenal experiences – is contradictory. Therefore, the absent qualia hypothesis is not only false: it is conceptually or logically false, and thus, its negation is logically true. The thesis supported by Tye can be seen as equivalent to the claim that conscious phenomenal experiences logically supervenes on functional organization:

20" Zuboff (1994, p. 183)

21" Zuboff (1994, p. 190)

> Necessarily, any system that functionally duplicates me is phenomenally conscious. The absent qualia hypothesis, therefore, is false even on its weakest interpretation.[22]

Cuda (1985) offers one of the first versions of the replacement argument in a paper suggestively titled *Against Neural Chauvinism*. Cuda's particular aims are, in part, to show that Searle's answer to the Brain Simulator reply, which has been discussed in the preceding chapter, is mistaken. Cuda interprets Searle as claiming that a necessary condition that a physical system needs to satisfy in order to possess mental properties is to be constructed of the right materials. Remember that Searle's position with respect to consciousness is that it is, first and foremost, a biological process, whose nature is not different from phenomena like digestion or photosynthesis, and as such, it seems unlikely that consciousness can be generated by entities whose physiochemical constitution differs greatly from the one of organic beings. Cuda, nonetheless, argues that Searle's position is chauvinistic, and presents his argument in favour of the claim that a system that is functionally equivalent to a human being at a neural level is a sufficient condition for having conscious states: "... functional equivalence to a human at a very fine level, is a sufficient condition for an organism to have conscious states."[23]

Thesis (SS) can be associated to the thesis of Multiple Realizability, which is the claim that mentality – and in particular, conscious experiences – can be realized by entities made out from very different materials. That a physical system is able to have conscious phenomenal states, like feeling pain, perceiving red qualia or experiencing sexual arousal, does not depend on it having such and such physical properties, for instance, the property of being composed of organic matter. Multiple Realizability can be seen as a consequence of thesis (SS): if a sufficient condition for mentality is the implementation of the right functional organization, then it seems that the physical nature of the different parts that form an entity is irrelevant, as long as these parts work correctly and perform the adequate function inside the system to which they belong. In (1994), Kirk offers a version of the replacement argument which explicitly endorses a form of the thesis of Multiple

22" Tye (2006, p. 159)

23" Cuda (1985, p. 124)

Realizability, and which is called by Kirk the Swiss Cheese principle (following a well-known example originally presented by Putnam). Kirk presents this principle as follows:

> … a thing's composition – what materials it is made of – has no essential bearings on (a) whether or not it has a mental life; (b) what mental states it has, if any; (c) what, if anything, it is like to be it. This is not to say that any materials whatever could be put together to make a mind. It might indeed prove impossible to make a mind out of cheese. The point is that the materials don't matter provided they do the right things, whatever those things might be. [24]

Note that Kirk's main point – in spite of emphasizing the role of the Swiss Cheese principle – is not to establish whether entities made out from any possible materials are capable of having conscious states. That must be seen as a secondary problem. The main issue is whether the pattern of interaction among the several parts of a system – its functional organization – gives rise to these conscious states. How can we understand, then, Kirk's claim that it might be impossible to build conscious beings from certain materials, like cheese, or perhaps water, wood or gin? The problem is not that there is something intrinsically wrong with some materials, but that they might be incapable, for instance, of instantiating the right functional organization with enough reliability, or even with the required speed. Consider the following analogy: to build a machine made out of soap, toothpicks and paper that exhibits the same computational capacities of a modern PC will be extremely challenging, maybe technically impossible. Some of the reasons are that, for instance, these materials do not have the same physical resistance of the materials from which a real PC is made. Also, they are much more prone to suffer undesirable alterations when they are exposed, for example, to the same changes of temperature or pressure that might affect the components of a PC. All this may impair the way these materials work and compromise the general reliability of the entire system. Moreover, and even accepting that such a machine could be built, it might not work with the same speed of a PC: since the physical resistance of the materials is not the same, they may break when they achieve certain velocity inside the system. Consequently, the machine might not interact with the

24" Kirk, (1994, p. 90)

environment in the same way that a PC: some external inputs would be too fast for it to process them.

As we have seen, Kirk rejects the use of the notion of supervenience in the formulation of the Strict Implication thesis. However, he argues that the Swiss Cheese principle implies the truth of this thesis. He reasons as follows: suppose that in the actual world, there are some non-physical entities that have a crucial causal role in the generation of mentality. But if the Swiss Cheese principle is true, then these crucial causal roles can also be performed by physical entities.

> An immediate consequence of the Swiss Cheese principle is that the Strict Implication thesis might be true. For even if it turned out that in the actual world certain non-physical items were involved in mental interactions, [...] ... the Swiss Cheese principle assures us that those same causal roles could have been performed by physical items instead. From the point of view of our interest in the phenomena of raw feeling, therefore, it doesn't matter whether mental interactions happen to involve non-physical items.[25]

(b) The level of functional organization

The notion of functional organization plays a fundamental role in the description of the general aims of the replacement argument, as well as in the way that the different replacement scenarios of the thought experiment are constructed. Roughly, the notion of functional organization can be understood as the abstract pattern of causal interaction that exists between the different parts of a system, and maybe also to the way that these parts interact with the inputs and outputs received and produced by it. In general, a functional organization F can be determined by specifying the following elements: (a) a certain number of abstract elements or parts; (b) for each of these elements, a number of possible states, and (c) a system constituted by dependency relations, which establishes how the state of each of these elements is determined by the previous states of all the other elements

25 " Kirk (1994, p. 105)

of the system and by the inputs received by it, and also which inputs are produced by the system depending on the previous states of its components.[26]

It is important to notice that the notion of functional organization describes an abstract entity. More precisely, it describes an abstract entity that can be implemented by a physical system. Most versions of the replacement thought experiment assume that physical systems – and in particular, organic brains of conscious beings – implement a determinate functional organization. However, there are some points that need to be clarified. First, it is very important to notice that most physical system – and this also applies to organic brains – implement more than one functional organization. This depends on how the different parts of the physical system are individuated, and how the states of these different parts are conceived. Second, authors like Searle (1994) have claimed that most physical systems not only implement more than a single functional organization, but that they also implement most of them. Also, an important assumption in this thesis is that the conditions that a physical system needs to satisfy in order to have a determinate functional organization are analogous to the conditions it needs to implement a certain computation. The relation between these conditions – not assumed by all versions of the replacement argument – is important insofar as one of the most discussed objections against functionalism (according to which there are no objective conditions for implementing a determinate computation) also affects the idea that physical systems implement a functional organization. Due to the fundamental role the notion of functional organization plays in the replacement thought experiment, it is essential to clarify the conditions under which a physical system implements a determinate functional organization.

In (1996) Chalmers suggests that the conditions under which a system implements a Combinatorial State Automaton are analogous to the conditions under which a system implements a certain functional organization. Chalmers suggestion is that this informal description of a functional organization can be formally described with the help of the notion of a CSA. The first step to understand the notion of a CSA is to deal with a simpler notion; the notion of a Finite State Automaton (which can be understood as a special case of

26" Cf. Chalmers, (1996, p. 247)

a CSA). A Finite State Automaton (FSA) is a mathematical abstract device, and as such, it is not necessarily designed for being implemented by a physical system. Briefly, it can be defined by the following structure:

$[\Sigma, \Gamma, S, s_0, f, \omega]$

The elements of this structure are defined as follows: Σ is the input alphabet, which is given by a finite set of symbols $\{i_1, \ldots i_n\}$. The output alphabet, Γ, is composed by a finite set of symbols $\{o_1, \ldots o_n\}$. S is a finite, nonempty set of formal states $\{s_1, \ldots s_n\}$. There is also an element of set S, s_0, which is known as the *initial state*. A state transition function, *f*, is defined as f: S x Σ → S. Finally, ω is the output function, and may depend on both a state and an input (ω: S x Σ → Γ), or only on a state (ω: S → Γ).

The framework of a FSA will serve as a basis for defining a CSA. In the case of a FSA, sets Σ, Γ and S, which correspond respectively to the set of inputs, outputs and internal states, are monadic: they are composed by single elements. In contrast, in a CSA these sets are structurally complex, and can be formally represented as follows:

$\Sigma = \{[i^1_1, \ldots i^k_n], \ldots [i^1_1, \ldots i^k_n]\}$

$S = \{[s^1_1, \ldots s^i_n], \ldots [s^1_1, \ldots s^k_n]\}$

$\Gamma = \{[o^1_1, \ldots o^i_n], \ldots [o^1_1, \ldots o^k_n]\}$

Another difference is that the transition rules of a CSA can be determined by specifying, for each element of set S, a function that determines how the next state depends on the-input vector and the previous state-vector, and also for each element of the output vector. These transition rules can be defined with the following notation:

$([i^1_1, \ldots i^k_n], [s^1_1, \ldots s^i_n]) \rightarrow ([s''^1_1, \ldots s''^k_n], [o^1_1, \ldots o^i_n])$

This expression can be read as follows: if the CSA is in vector-state [s^1_1, ... s^i_n]) receiving input-vector [i^1_1, ... i^k_n], it will transit into vector-state [s''^1_1, ... s''^k_n] and will produce output [o^1_1, ... o^i_n]). The conditions under which a system implements a CSA can be precisely formulated as follows:

> A physical system implements a given CSA if there is a decomposition of its internal states into substates [s_1, s_2, … s_n]and a mapping *f* from these substates onto corresponding formal states S^j of the CSA, along with similar mappings for inputs and outputs, such that: for every formal state transition ([I^1, … I^k], [S^1, ... S^n]) → ([S''^1, ... S''^k], [O^1, ... O^j]) of the CSA, if the system is in internal state [s^1, … s^n] and receiving input [i^1, ... i^n] such that the physical states and inputs map to the formal states and inputs, this causes it to enter an internal state and produce an output that map appropriately to the required formal state and output.[27]

Let's see how the functional organization of Paul´s brain at a neural level can be abstracted into a CSA. The way for doing this is to identify a number of basic elements – in this case, organic neurons – and to stipulate that to each state vector of the CSA corresponds one of these basic elements. Also, it is necessary to stipulate that the causal interactions among these organic neurons correspond to the formal transition rules of the CSA. This can be better understood if we consider the replacement process that generates Paul´s functional isomorph at a neural level, F-Paul. Suppose that F is the functional organization instantiated by Paul McCartney´s brain at the neural level. According to the framework provided by a Combinatorial State Automaton, functional organization F can be abstracted into a CSA_F. Any system that implements CSA_F according to the conditions suggested by Chalmers will have F as its functional organization. For each organic neuron inside Paul´s brain, there will be a corresponding vector of CSA_F that determines how the internal state of this neuron depends on the state of other neurons. The replacement procedure starts by extracting one organic neuron from Paul´s brain and then installing in its former place an artificial neuron that duplicates the function of the organic neuron by implementing the corresponding vector. This implementation consists in following a series of formal transition rules that determine the way in which the artificial neuron, after receiving a stimulus from the presynaptic neurons in the neighbourhood, will in turn send a

27 " Chalmers (1996b, p. 325)

corresponding stimulus to the postsynaptic neurons. This replacement procedure will continue until for each organic neuron in Paul´s brain there is a corresponding artificial neuron. In this way, the system composed by these artificial neurons – F-Paul´s artificial brain – shares the same functional organization of Paul´s organic brain by implementing CSA_F. This framework will allow us to formulate a modified version of theses (NSS) and (LSS):

(NSS_M) There is a set *C* of Combinatorial State Automata such that the property of a system *P* of generating conscious phenomenal experiences *naturally* supervenes on the property of *P* of instantiating a member of set *C*.

(LSS_M) There is a set *A* of Combinatorial State Automata such that the property of a system *P* of generating conscious phenomenal experiences *logically* supervenes on the property of *P* of instantiating a member of set *C*.

Most versions of the replacement thought experiment consider a level of functional organization in which the organic neurons of the brain of a conscious person are identified as the basic elements of this functional organization. This is the case of the thought experiments proposed by Cuda, Kirk and Chalmers. As we mentioned before, this strategy has a very good justification: neurons are the most basic elements of the brain, and as we mentioned in the introduction, there is excellent empirical evidence in favour of the claim that the brain is the organ directly responsible for the generation of conscious phenomena experiences. Of course, the brain of Paul can implement another functional organization if we identify another set of basic elements instead of organic neurons. The replacement thought experiment described in Zuboff (1994) is a notorious example: Zuboff considers a level of functional organization in which regions or parts of the brain associated to certain mental function are the basic elements of replacement. The replacement process imagined by Zuboff is similar to the one just described; although in this case these basic elements are replaced by devices or gadgets that, according to Zuboff, duplicate the causal effects of the original region.

(c) The features of the replacement entities

Another variable element of the replacement thought experiment corresponds to the entities that replace the basic elements of the brain (or in Tye´s case, the physical realizers of internal states). In the literature, these entities include tiny homunculi inside capsules that follow a set of instructions (Cuda), gadgets that duplicate the causal relations that organic regions of the brain have with respect to the whole system (Zuboff), or miniature computers that implement a certain vector state of a CSA (Chalmers). Zuboff even considers the possibility of replacing regions of the brain with an entity that, by pure chance, generates the same electrochemical output signals produced by the visual cortex. This entity might be an empty shell in which these signals are produced randomly. Of course, the way in which these signals are produced might be chaotic, but Zuboff propose to imagine that this shell produces the same output signals produced by the visual cortex for a certain amount of time, say, sixty seconds.

What is important is that the replacement preserves the same connections that the original organic part has with the rest of the brain. At the neural level, if the organic neuron produces an output signal in response to a certain stimulus, the replacement would be sensible to the same stimulus and will produce the same output. If the replacement is a tiny computer, it will be equipped with a device that produces the adequate electrical impulses, and perhaps with another device that stores and produces the adequate neurotransmitters, always following a certain program. If the replacement consists in a homunculus inside a capsule, it might follow a set of instructions that indicate the electrical and chemical signals that need to be produced when other electrical and chemical signals are received.

(d) The features of the replaced entities

In most versions of the replacement argument, neurons are the entities that are replaced by functionally equivalent systems. This strategy has a very good justification: neurons are the most basic elements of the brain, which is considered the organ responsible for mentality in general. Tye´s version, however, can be better conceived as a mutual

replacement, or interchange, of the internal states of two systems S and Sə that are functional duplicates. The internal states of these two beings are conceived also as functionally isomorphic (in a sense that will be explained in the first section of the fourth chapter of the thesis), but differ in that, while the states of being S have phenomenal properties, the states of Səare phenomenally inert. Tye´s strategy is to assume, as a *reductio* hypothesis, that the absent qualia hypothesis implies the logical conceivability of a being like Sə This assumption, however, contradicts with a principle that is conceived by Tye as necessary, and thus, Tye concludes that a being like Səis not logically conceivable, and thus, that the absent qualia hypothesis is logically false.

Conclusions

In this chapter, I had three main objectives. The first was to present the thought experiments proposed by Block and Searle. The objective of these thought experiments is to show that the duplication of the functional organization of the brain of a conscious person, at least at the neural level, is not a sufficient condition for the generation of mentality, and in particular, is not a sufficient condition for generating conscious phenomenal experiences. The versions of the replacement thought experiment that will be presented in this thesis can be understood as showing that, in spite of the bizarre nature of the systems described by Block and Searle, systems that duplicate the functional organization of the brain at the adequate level are capable of generating conscious phenomenal experiences.

My second objective was to argue that the formulation of the theses supported by the several versions of the replacement thought experiment can be expressed with the help of the notion of strong supervenience formulated in terms of quantifiers over possible worlds. The first was thesis (NSS): there is a set F of functional organizations such that the property of a system P of generating conscious phenomenal experiences *naturally* supervenes on P instantiating a member of set F. The argument presented by Chalmers in (1996) is explicitly given as supporting thesis (NSS). The second thesis was (LSS): there is a set F of functional organizations such that the property of a system P of generating

conscious phenomenal experiences *logically* supervenes on P instantiating a member of set F. Authors that propose versions of the replacement thought experiment that support thesis (LSS) are Zuboff (1994) and Tye (2006). Finally, although the version of the replacement argument proposed by Kirk in (1994) does not explicitly support a supervenience claim, it can be understood as supporting a thesis with a similar modal character. His version is presented explicitly as supporting a version of the thesis of Multiple Realizability, called by him the Swiss Cheese Principle. Kirk argues that this principle supports what he calls the Strict Implication Thesis, according to which the set of sentences P that correctly describe the past, present and future of the whole universe logically imply a set Q of sentences that ascribe mental states to the individuals of this universe.

Finally, in order to provide a more accurate characterization of the several versions of the replacement thought experiment, I identified some of the main variable elements of the replacement scenario described in them. These variable elements are (a) the modal character of the functionalist thesis supported by the different versions of the thought experiment, (b) The level of functional organization at play, (c) The features of the replacement entities, and finally, (d) The features of the replaced entities.

Chapter 2
Two initial versions of the replacement thought experiment

Introduction

In this chapter, I will present and evaluate two initial versions of the replacement thought experiment, as well as associated arguments derived from it whose objective is to show that thesis (NSS) is true: there is a set F of functional organizations such that the property of a system P of generating conscious phenomenal states supervenes on the property of P of instantiating a member of set F. Remember that if thesis (NSS) is true, the absent qualia hypothesis is at least empirically false: in a possible world whose physical laws are the same ones as in the actual world, it is not the case that functional isomorphs differ in that only one of them is capable of generating conscious phenomenal experiences.

In the first section of this chapter, I will discuss an initial replacement strategy that does not involve the functional duplication of the brain of a conscious being, but a simulation at the microphysical level. As we will see, this strategy preserves the conscious phenomenal experiences of the subject. However, this strategy cannot be used for arguing in favour of thesis (NSS). The reason is that this strategy is conceived as duplicating the microphysical structure of a conscious being, and not its functional properties.

In the second section of this chapter, I will present a first version of the replacement thought experiment based on the strategy proposed by Zuboff in (1994).[28] According to the framework presented in the last chapter of this thesis, this version describes a replacement scenario in which the functional organization of the brain of the subject is relatively less fine-grained than the neural level. In this case, the base properties are defined as the

28 "Although Zuboff's thought is explicitly presented as supporting the claim that functional organization logically determines phenomenal consciousness, my aim in this chapter is to determine whether a similar version of the thought experiment that considers the same functional organization (that is, the functional organization of the brain at the regions-of-the-brain level) is capable of supporting the claim that phenomenal consciousness naturally supervenes on functional organization.

property of instantiating the same functional organization of the brain at the *regions-of-the-brain* level.

In the third section of this chapter, I will present some problems faced by the *regions-of-the-brain* version of the replacement thought experiment, mainly related to the properties of the replacement entities proposed in this version of the thought experiment. If these entities are defined as duplicating all the causal relations that the original region had with respect to the rest of the brain, the replacement process might succeed in creating an entity capable of experiencing conscious phenomenal states. However, I will argue that it is at least doubtful that a gadget that does not share the same physicochemical composition of the original brain region can duplicate all the causal relations that this original part had with respect to the rest of the brain. In order to duplicate these casual relations, the gadget would have to be almost identical to the original organic region. But if this is so, this version of the replacement thought experiment would not show that the capacity of generating conscious phenomenal experiences supervenes on the property of instantiating a certain functional organisation. The reason, of course, is that the physicochemical composition of the gadget might have a crucial role in the generation of these experiences.

In the fourth section of this chapter, I will present a version of the replacement thought experiment that considers the functional organization of the brain at a neural level, which is mainly based on the version presented by Chalmers in (1996 and 2006). My objective is to determine whether this version supports at least thesis (NSS): there is a set F of functional organizations such that the property of a system of generating conscious phenomenal experiences naturally supervenes on P instantiating a member of set F. This version of the thought experiment adopts a sort of *reductio* strategy (although it is important to notice that it is not a genuine *reductio*, since its objective is not to show that the assumption of the absent qualia hypothesis leads to a logical contradiction). Thus, the "*reductio*" hypothesis will be formulated as follows: it is naturally possible that a system that duplicates the functional organization of the brain of a conscious person at a neural level lacks the capacity of generating conscious phenomenal experiences. The strategy will be to consider a replacement process in which the organic neurons of the brain of a

conscious person – Paul McCartney – are replaced by artificial neurons. As the replacement process is done neuron for neuron, it generates a sequence of beings that preserve the functional organization of Paul´s brain at the neural level. Finally, in the fifth section of this chapter, I will discuss some objections related to the claim that a system can duplicate the input-output behaviour of an organic neuron.

Section 1
The microphysical simulation of the brain of a conscious being

Before presenting and evaluating the regions-of-the-brain version and the neural version of the replacement thought experiment, I want to briefly discuss a different replacement strategy proposed originally by Block in (2007b). This replacement strategy depicts a situation in which the most basic components of a conscious being – namely, subatomic particles – are replaced by other entities, while the ability of this being of experiencing conscious phenomenal experiences is preserved. The aim of the replacement strategy I will present in this section, as it was originally proposed by Block, is not to show that the preservation of the functional properties of a physical system is sufficient for preserving its mental capacities. Block designs this strategy in order to argue against what he considers an *ad hoc* approach for objecting to the Chinese Nation thought experiment. According to this strategy, one can stipulate that if systems A and B differ in that A contains elements with functional organizations that are characteristic of beings capable of experiencing conscious phenomenal states, while B does not, then A and B cannot be functionally equivalent. In (1967), Putnam argues that the claim that "being in pain is a functional state of an organism" can be defined as follows:

> (1) All organisms capable of feeling pain are Probabilistic Automata.
> (2) Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e. being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).
> (3) No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in (2).
> (4) For every Description of the kind referred to in (2), there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in

that subset.[29]

According to clause (3), if an organism is able to experience pain, then its constituent parts cannot be described as being capable of feeling pain. If we accept this stipulation, then the Chinese Nation cannot be functionally equivalent to the brain of a conscious person. The reason is that the Chinese Nation includes parts that are composed of Chinese people, able to experiencing conscious phenomenal states, while the parts of the brain of that conscious person are unable, by themselves, to have these experiences. Now, the problem seems to be that, in the case of the Chinese Nation, the system acquires the relevant functional organization in virtue of the fact that conscious beings play a certain fundamental role inside the system. Thus, the *ad hoc* proposal can be formulated as follows: a system capable of having conscious phenomenal experiences in virtue of instantiating a certain functional organization cannot possess, among its constituent parts, conscious beings that play a certain crucial role in giving the system its functional organization.

The strategy proposed by Block can be interpreted as describing a system that has a certain functional organization in virtue of the role played by this sort of conscious beings, and at the same time is capable of having conscious phenomenal experiences. Imagine that in some unknown and probably very different region of the universe, there are extremely small, intelligent beings that are composed by matter that is infinitely divisible and is also very different from the matter we know (for ease of exposition, we can call it here "h-matter"). Suppose that these intelligent beings, at some point in their history, travel through the universe and discover the existence of our type of matter (call it "r-matter"). For some reason known only to them, they decide to construct flying machines that resemble all the elementary particles that compose r-matter. They build, for instance, machines that resemble protons and neutrons. Powerful engines attached to these machines bind them together in order to simulate the atomic nucleus. Other machines that move around this nucleus simulate the behaviour of electrons. These flying machines are constructed in such a way that they mimic the subatomic processes that generate the basic properties of r-matter. These intelligent beings have at their disposition an extremely large amount of h-

29 " Putnam, (1967, p. 434)

matter for constructing these flying machines, and soon they construct a huge amount of "substances" made from h-matter that resemble, at the subatomic level, the matter we know. They start by constructing simulations of basic elements, like hydrogen and helium. Little by little, the amount of simulated matter increases and, after some millions of years, there is a region of the universe in which stars and planets made from this simulated matter can be found.

Imagine now that a group of astronauts travel to that region of the universe, where they find a planet that resembles the conditions on Earth. When they arrive, they discover that can breathe the "air" in its atmosphere and drink the "water" on its rivers. They also discover things that grow in the surface of the planet, very similar to the plants and vegetables that can be found at Earth, and find that they are edible. Convinced that the environment in that planet is adequate for supporting human life, they decide to establish a colony. After some years of living there, a very interesting phenomenon occurs: the body of the astronauts becomes composed of this artificial matter. The reason, of course, is that the molecules of their bodies are gradually interchanged with the "molecules" of the environment.

Assume that the astronauts travel back to Earth after living for several years in that part of the universe. When they arrive, a group of physicians – perhaps after putting them in quarantine – deicide to evaluate them medically. The physicians, however, do not find anything odd. The astronauts still have hearts, livers, kidneys and brains that work in more or less the same way as the organs of the people that remained at Earth. Importantly, their brains preserved the structures to which a role in the generation of mentality is commonly attributed. Inside the heads of the astronauts, there are tiny structures that are indistinguishable from real neurons. The physicians notice that these structures form large networks in which electrochemical signals are interchanged through connections identical to synapses. The generation of these electrochemical signals is also identical to the way in which they are generated by the neurons of normal people: there is a sodium-potassium pump that regulates the interchange of ions through the membranes of these structures. Of course, some people might reject the idea that that these structures could be called "hearts",

"kidneys", "brains", "neurons", or "synapses". By hypothesis, these internal structures are not composed of organic matter as we know it. But at an extremely fine-grained level, they work in exactly the same way as the organic, biological internal structures of a normal person. All the physical phenomena that might contribute to the generation of mentality and of conscious phenomenal experiences has a corresponding correlate in the simulation just described.

> The basic electrochemical mechanisms by which the synapse operates are now fairly well understood. As far as is known, changes that do not affect these electrochemical mechanisms do not affect the operation of the brain, and do not affect mentality. The electrochemical mechanisms in your synapses would be unaffected by the change in your matter.[30]

Is this sufficient for granting that the astronauts still preserve the capacity of having conscious phenomenal experiences? I can think on two different objections against the claim that the astronauts did not preserve their mental capacities after their molecules in their brains and bodies were replaced by molecules composed of h-matter. The first is that h-matter might introduce an undetectable but detrimental effect in the mental life of the astronauts, in such a way that when their physical composition changes, h-matter progressively destroys their mental properties. After all, h-matter is completely different from the matter we know and we are composed of, and we simply cannot deny the possibility that one of its properties was that it makes impossible the presence of any mental property. But how these detrimental effects would be manifested in the brains of the astronauts? By hypothesis, this simulated matter behaves exactly like real matter. If h-matter would have this harmful property, it would not be manifested in the way in which the simulated particles behave, and particularly, it could not then affect the way in which the synapses of the astronauts work.

The second objection goes as follows: perhaps the level of simulation described is not the most adequate. It might be argued that the generation of mental phenomena is related in some way to the effects produced by subatomic particles that were not simulated

30 " Block, (2007b, p. 75)

by the machines created by these tiny beings. Protons and neutrons belong to a larger family of subatomic particles, the hadrons, which are in turn composed by quarks. There are, perhaps, quark-related phenomena that the simulation described cannot represent, simply because the level of simulation chosen is not as fine grained as the level of quarks, and these phenomena might have an essential role in the generation of mentality. However, it is not difficult to imagine a further modification in the simulation process in which the tiny beings decide to simulate these more basic subatomic particles. Instead of protons and neutrons, they can create machines designed to simulate the behaviour of quarks. By hypothesis, the matter used for creating these machines is infinitely divisible, so there does not seem to be any problem for simulating r-matter at the level required.

There is, of course, a difference between the cases presented by Block´s Chinese Nation and Searle´s brain simulator and the simulation of subatomic particles just described. Notice that, in contrast with these former cases, the objective of the tiny beings is not to build a system that implements the functional structure of conscious people, but to simulate their microphysics. Of course, the progressive replacement of the molecules of the bodies of the astronauts does not modify the way in which their brains process information. As it has been mentioned, no medical procedure can detect a difference between them and the people that remained on Earth. The same neurophysiological theories that apply to normal people apply to these astronauts. In contrast, these theories do not apply to systems like the ones exemplified by Block´s Chinese Nation and Searle´s brain simulator.

> There is one very noticeable difference between the elementary-particle-people example and the earlier homunculus examples. In the former, the change in you as you become homunculus-infested is not one that makes any difference to your psychological processing (i.e., information processing) or neurological processing but only to your microphysics. No techniques proper to human psychology or neurophysiology would reveal any difference in you. However, the homunculi-headed simulations [...] are not things to which neurophysiological theories true of us apply [...] This difference suggests that our intuitions are in part controlled by the not unreasonable view that our mental states depend on our having the psychology and/or neurology we have. [31]

31 " Block, (2007b, p. 76)

Section 2

The *regions-of-the-brain* version of the replacement thought experiment

In (1994), Zuboff offers a version of the replacement thought experiment. Among the several functional organizations instantiated by the brain, this version of the thought experiment considers what I will call here the parts-of-the-brain level of functional organization. We have very good empirical information in favour of the claim that some regions of the brain are associated to certain mental properties. For instance, we know that the right and left visual cortices, which are located, respectively, in the left and right hemispheres of the brain, are the parts of the cerebral cortex that are responsible for processing visual information. The left hemisphere visual cortex processes signals from the right visual field, while the right hemisphere visual cortex processes signals from the left visual field. Both visual cortices receive, in turn, signals generated by the lateral geniculate nucleus, which is the part of the brain that receives visual information directly from the retina of the eye.

The replacement scenario depicted by this version of the thought experiment assumes the possibility of building non-organic devices capable of duplicating the input-output behaviour of sections of the brain associated with a particular mental function. What these organic devices do is to replace their organic counterparts by interacting with the rest of the brain in exactly the same way. The process continues until none of the organic parts of the brain remain and a new system, which shares the same functional organization of the brain at the level just described but that is composed entirely by these artificial replacements, is created. The argument built on this replacement scenario is directed to show that the preservation of the functional organization of the original system in the new system preserves the mental properties (if any) of the former one.

Imagine that a group of highly competent neuroscientists replace Paul McCartney´s visual cortex with a gadget or device that, according to Zuboff, “will keep precisely the

same relationship with the rest of the brain that the replaced chunk had". What the precise character of this relationship is will be discussed later, but at least, we can be sure that this includes all electrochemical signals that the original visual cortex transmits to the neighbouring parts of the brain. Now, since the rest of the brain and in particular the regions concerned with the production of speech are causally affected by the gadget (lets call it GV) in the same way they would be affected by Paul McCartney´s original visual cortex, his linguistic behaviour (and in fact his entire behaviour) would not be different from the behaviour he would have exhibited if the replacement had not been performed. If he were in front of a flying lady carrying a bag of diamonds and somebody asked him "Hey, Paul, can you see that lady over there?" he would answer "Yes, I can" if he wanted to give a sincere response. His behaviour in general would be indistinguishable from a person having visual experiences. But – Zuboff argues – it would be absurd to think that he could exhibit such behaviour if he did not have these visual experiences. Therefore, such experiences must have been preserved after the replacement of Paul´s visual cortex by a functionally equivalent system.

> But think about this: it would be absurd for us thus to be assured that you would go on behaving and speaking the same after the replacement if it were possible for us to think that your experience might have been different from what it would have been with the chunk of brain unreplaced. If the replacement by wires and transistors in that part of brain activity could have made you see or hear or feel or think any differently, how could we have the assurance our stipulation must give us that you would not do or say anything different? (Anyone who is not startled by this step in the argument is probably not understanding it.) A gadget that saves the pattern of mental functioning must, surprisingly, therein have saved the experience too.[32]

Thus, the moral that Zuboff draws from this argument is that the function performed by the original organic region of the brain will be preserved after the replacement. In particular, if we replace the visual cortex of Paul McCartney by a device that implements the same input-output causal relationships, the phenomenal character of his visual experiences will not be modified. The following is a more precise reconstruction of Zuboff's argument:

32 " Zuboff, (1994, p. 183)

(Premise 1) There is a gadget GV that duplicates the same causal relationship that Paul McCartney´s visual cortex has with respect to the other regions of his brain.

(Premise 2) Since GV preserves the same causal relationships that the original Visual Cortex had with respect to the rest of the brain, the regions responsible for the generation of Paul McCartney's linguistic behaviour will be affected exactly in the same way by GV.

(Premise 3) Since the regions of the brain responsible for the generation of Paul McCartney's linguistic behaviour will be affected in the same way by gadget GV, Paul McCartney's linguistic behaviour will be exactly the same.

(Premise 4) It would be absurd to think that Paul McCartney preserved his linguistic behaviour and, at the same time, affirming that the replacement eliminated his visual experiences.

(Conclusion) Paul McCartney's visual experiences were not modified by the replacement of his visual cortex by gadget GV.

As it is indicated in premise (4), there is something that is *prima facie* problematic in saying that a functional isomorph can behave as if it had visual experiences but without having them at all. However, this does not seem to be completely absurd. The fact that the behaviour of Paul McCartney – and specially, his linguistic behaviour – is exactly the same after the replacement does not seem to show, by itself, that his visual experiences are preserved when his organic visual cortex is replaced by gadget GV. Just because Paul McCartney's behaviour is consistent with him having visual experiences we cannot conclude that he, in fact, has them. After all, this is precisely what happens with complete functional zombies: their behaviour is consistent with the possession of conscious mental states, but they completely lack mental life. Cannot we say that after the replacement of the Visual Cortex, Paul McCartney is something like a visual zombie that deep inside is just like a blind subject but that exhibits the behaviour of a normal human being?

In spite of these considerations, it is important to remember that at this stage of the replacement process, Paul McCartney is by no means a full zombie. Even if his visual experiences are eliminated after the replacement, nothing in the envisaged scenario prevents him for forming new beliefs about his current behaviour. But if this is possible, it

would be odd to think that Paul McCartney can form new beliefs concerning a behaviour that is consistent with him having visual experiences, and at the same time not having these experiences at all. This is an initial description of the incompatibility noticed by the proponents of these replacement scenarios: the one that exists between the subject's lack of conscious phenomenal experiences (visual, auditory, olfactory, etc.) and the fact that the subject seems to be able to generate conscious intentional states about these experiences and about a behaviour that is consistent with its presence. However, Zuboff´s argument seems to rely only in the fact that the linguistic behaviour of the subject will be consistent with the presence of visual experiences, a fact that, as I have argued, does not suffice for showing that these experiences are preserved after the replacement process. In section (4) of this chapter I will present a different version of the replacement thought experiment and associate arguments whose objective is to show that there is an incompatibility between the fact that the subject preserves not only the same linguistic behaviour after the replacement, but also between the fact that his beliefs and other cognitive states related to his conscious phenomenal experiences are preserved. Meanwhile, I want to discuss some further objections concerning the parts-of-the-brain version of the replacement thought experiment.

Section 3

Objections to the *regions-of-the-brain* version

A further objection to the parts-of-the-brain version of the replacement scenario can start as follows: the proposal that a physical system can duplicate the input-output function of an organic region of the brain is extremely unclear. The replacement of the visual cortex, for instance, by one of these systems, may be simply to play with the brain, to mess around with it in such a way that any result would be completely unpredictable. Most probably, the consequence of such replacement would simply be a general malfunction related to the properties of processing visual information and of generating conscious visual experiences, similar to the destruction of the original organic part. Even if some of the functions usually attributed to the visual cortex are preserved after the replacement, we simply could not know which ones: the number of variable elements that need to be considered to make an

adequate prediction is very high and there is no systematic way of knowing how the brain would react to this replacement.

In order to get a better understanding of this objection and the possible responses, it will be convenient to examine some of the proposals that explain what a physical system needs to do in order to duplicate a certain region of the brain. Let's start precisely with the way that Zuboff describes the details of this duplication. Zuboff imagines a physical system or gadget that, as he stipulates, duplicates exactly the same causal relationships that the visual cortex has with respect to other regions of the brain. The preservation of these causal relationships allows the gadget to preserve the visual experiences of the subject.

> Let's imagine that a chunk of your brain was to be replaced by a wire and transistor gadget that, as we shall just stipulate, will keep precisely the *same* causal relationship with the rest of the brain that the replaced chunk had. We can know, based merely on this stipulation of the sameness of the gadget's effects on the rest of the brain, that you will behave and speak exactly as you would have done if the circumstances were otherwise the same but no such replacement of a chunk of the brain had been made. For the parts of the brain responsible for speech and behavior must, according to the stipulation, be affected by the gadget in all ways as they would have been by the normal brain chunk.[33]

Note that a brain region may have a number of different causal relationships with the rest of the brain and the body. Some of its properties, like its weight and density, are determined by its physical composition, and these properties may have a particular causal effect on the rest of the brain. It is true that these properties may not be related to what is commonly understood as the proper function of the visual cortex – that is, the processing of visual information –, but as it will be argued, they may still be relevant. Imagine, for instance, that the builders of the gadget use in its construction components made from stainless steel. If the amount of this material in the gadget is high, its weight might be much larger than the weight of the original visual cortex. This may have unintended effects on the rest of the brain and affect the way it works. Also, another problem with a visual cortex made out from a high percentage of steel components is that it may react to magnetic fields in a different way from the original organ. Or perhaps the problem is that the gadget, in

33 " Zuboff, (1994, p. 183)

spite of duplicating all the causal effects of the visual cortex, also introduces some other unintended effects that impair the way in which the rest of the brain works. This might be so if the replacement gadget includes components made out from, perhaps, plutonium, which is a highly toxic element. Of course, someone may reply that this does not respect Zuboff's stipulation that the gadget should keep exactly the same causal relationships with the rest of the brain that the original visual cortex had. However, it is not clear then what materials can be used for constructing such device. Is it possible to build an artificial visual cortex that shares all and only its causal properties?

The point I want to make here is that these issues show that the materials from which the replacements are made are, after all, important. If the aim is to build a device with the same causal relationships that the visual cortex has with respect to the rest of the brain, it seems that the composition of the replacement device cannot differ excessively from the original part. In fact, it is possible to think that the only way for duplicating the function of the visual cortex without introducing any unintended effect on the rest of the brain and the body is to build a biologically equivalent visual cortex that shares not only the same functional organization but also its physicochemical composition. In other words, it might be possible that the only physical system able to duplicate the casual powers of the visual cortex is a biologically equivalent visual cortex.

Now, there does not seem to be any *a priori* difficulty in building an artificial gadget with these characteristics. It is true that its construction may be a serious technical challenge, and moreover, it may be something that cannot be achieved with our current technology. But this is clearly an empirical matter that cannot be settled here. Nonetheless, there is a more pressing issue. If the objective of the replacement thought experiment is to give support to the supervenience thesis – that is, the thesis that the property of a physical system P to have conscious phenomenal experiences supervenes on P implementing the right functional organization – then it is doubtful that this version of the replacement argument achieves this goal. The reason is precisely that the physiochemical composition of the replacements would need to be very similar, if not identical, to the organic visual cortex. Only these systems would implement adequately the right functional organization. But if

the only system that can duplicate the causal effects that the visual cortex has with respect to the rest of the brain is an organically equivalent device, then it is doubtful that the thought experiment shows that the functional organization of the system determines its capacity of generating conscious phenomenal experiences. We simply cannot rule out that a property related to the physiochemical composition of the gadget is responsible of generating these experiences.

In spite of these considerations, I think there are good reasons for dismissing these worries against the *regions-of-the-brain* version of the replacement scenario. Note first that a gadget like the one proposed for replacing the visual cortex can be seen as a device much similar to those designed for replacing damaged bodily organs. Artificial organs can be built for restoring numerous functions previously lost, or even absent from birth. Think for instance, of artificial devices designed for replacing amputated limbs or broken bones, which in some circumstances are very effective and allow the patient to recover several of his previously lost abilities. In some other cases, artificial organs can also provide artificial life support for patients during dangerous surgeries, for instance, when the patient is awaiting a heart transplant. Note that, none of these artificial devices need to share all the causal relationships that the replaced organs have with the rest of the body of the individual. Moreover, they do not need to duplicate all these relationships in order to perform adequately the function of these original organs. Of course, it can still be argued that the physiochemical constitution of the gadget matters. There are some materials that are obviously inadequate for constructing the proposed gadget, like those capable of damaging the brain or the body. Notoriously, this sometimes happens with real, artificial organ replacements. For instance, artificial hips made from cobalt may have certain harmful effects on patients. When the metallic parts of the artificial joints grind against each other, they release microscopic fragments of cobalt in the patient's body, which are toxic. On occasions, the excess of cobalt is naturally expurgated, but sometimes it can accumulate in the body. However, the fact that some artificial organs have unintended, and even harmful effects on the body does not show that they do not perform their intended functions adequately. In the same way, a gadget replacing the visual cortex may include among its components some made from, say, plutonium, which is also toxic and extremely harmful.

Even though, and apart from having these negative consequences, the gadget may be perfectly capable of duplicating the function of the visual cortex, in the same way that a hip made from cobalt may duplicate the function of the biological organ.

How can then a gadget duplicate the relevant function of the visual cortex without sharing all the causal relationships that it has with respect to the rest of the brain? What are the properties that the gadget needs to have in order to duplicate the function of the visual cortex? The visual cortex also has a set of properties that have a direct role in the generation of visual experiences. These are its neural properties. It can be stipulated that the device that replaces the visual cortex duplicates only these properties: what the device does is to receive electrochemical signals through the connections that the original visual cortex had with the rest of the brain, and in turn produces a corresponding output electrochemical signal. In fact, this is the way in which Zuboff describes the replacement process:

> Among the many properties of the visual cortex, neural, chemical, and computational (and let me also mention the imagined property of generating epiphenomena), there is the functional property of the visual cortex, its purely extrinsic property of causing a particular pattern of effects in the various mental functions. But the visual cortex possesses this extrinsic functional character only because its intrinsic neural and other properties have combined to produce the required pattern of external effects.[34]

Most versions of the replacement argument share the assumption that, in biological brains, what matters for the generation of mentality, and in particular, for visual experiences, is the neural pattern of the visual cortex. The visual cortex has several connections to the rest of the brain. For instance, it receives electrochemical impulses via a collection of axons from the Lateral Geniculate Nucleus, which in turn is the region of the brain that receives visual information from the retina. A replacing gadget may need to be connected to the same axons and to be able to receive and process these electrochemical impulses in order to produce the adequate output. But note again that it is not necessary to stipulate that the gadget duplicates all the causal relationships that the visual cortex has with the rest of the brain, like those related to its weight and physical composition. In this

34 " Zuboff, (1994, p. 185)

way, it might be possible to avoid objections concerning the idea that the only physical device that shares all the causal relations of the visual cortex is, precisely, the visual cortex.

There is, however, a further worry concerning the *regions-of-the-brain* version of the replacement thought experiment. Notice that, in some cases, the removal of certain parts of the brain does not cause the complete loss of the mental functions associated with them. In (2012), Majorek mentions the case of a patient that suffered from Rasmussen´s encephalitis, which is an infrequent neurological disease that affects a single cerebral hemisphere and that may generate the loss of motor and speech abilities, paralysis, encephalitis and dementia. The cerebral hemisphere affected by this illness was surgically removed. The physicians considered that the risk of total paralysis of the body and the complete loss of linguistic abilities was preferable to the danger of a more acute case of encephalitis. However, the patient – a girl that underwent the operation at the age of three years – was proficient in English and Dutch at the age of seven, with only minor problems related to the movement of her left arm and leg.[35] There is no definitive explanation of how the linguistic abilities of the patient were preserved after the removal of the brain hemisphere. However, a possible explanation lies in what has been called the *plasticity* of the brain. Plasticity is a property of the brain that consists in the increase or decrease in the number of brain cells, as well as the modification of the synapses. In (2005), Pascual-Leone *et al* characterize the plasticity of the brain as follows: "... changes in the input of any neural system, or in the targets or demands of its efferent connections, lead to system reorganization that might be demonstrable at the level of behavior, anatomy, and physiology and down to the cellular and molecular levels."[36] A consequence of the plasticity of the brain is that the location of the brain activity linked to a certain mental function can be modified. Now, it might be argued that the preservation of the visual experiences of Paul after the replacement depended not on the characteristics of the replacement gadget, but simply because the brain processes that generated visual experiences and that were located in his visual cortex were relocated to a different place in his brain.

35 " Cf. Majorek (2012, pp. 123-125)

36 " Pascual-Leone et al., (2005, pp. 378-379)

I do not think that this is a definitive objection against the *regions-of-the-brain* version of the replacement thought experiment, but I will not press this point further. In the following section of this chapter I will present a version of the replacement thought experiment that considers a more fine-grained level of functional organization, namely, the functional organization of the brain of a conscious person at the neural level.

Section 3

The neural version of the replacement thought experiment

Remember the replacement scenario already mentioned in the introduction of the thesis. After a terrible accident, a group of scientists suggest Paul McCartney an operation that might save him from a neural disease that threatens his life. The operation consists in replacing the organic neurons in his brain by artificial neurons that will preserve the same input-output behaviour of these organic neurons.

As the replacement process is done neuron by neuron, we can imagine a very large sequence of individuals between Paul McCartney and F-Paul that results from the replacement of all the neurons in his brain by artificial neurons. Let us say that the first member of the sequence is precisely Paul McCartney, since the number of neurons replaced in his original brain is zero, while the last member is F-Paul, whose artificial brain is entirely composed by artificial neurons. The mentioned sequence can be represented as follows:

$< J_0, J_1, J_2, , \ldots , J_n >$

Note first that each J_i that belongs to this sequence is functionally identical to Paul McCartney and to each other, or in other words, that the functional organization of each of the members of this sequence is exactly the same, so if there is a feature of Paul McCartney that depends on the functional organization of Paul McCartney´s brain, it must have been preserved.[37]

37 " Whether or not this replacement preserves the personal identity of Paul is something that may not be ruled out.

Most of the arguments that are based on replacements scenarios like the one just described proceed by adopting a *reductio* strategy: we assume that the absent qualia hypothesis is true. More precisely, we adopt the hypothesis that the last member of the sequence, F-Paul, lacks the capacity of having conscious phenomenal experiences. Remember that according to the discussion of the absent qualia hypothesis presented the first chapter of this thesis, it can be understood at least in the following two ways:

(1) Absent qualia are empirically possible.

(2) Absent qualia are logically possible.

The strategy I will adopt in this chapter is to argue, by examining a number of cases, that the *reductio* hypothesis is incompatible – or even inconsistent – with previous assumptions or with purported empirical knowledge concerning how the brain works. It is important to remember that some of these arguments proceed by arguing that the *reductio* hypothesis generates a genuine contradiction, as in the case of the versions offered by Zuboff and Tye. In the version offered by Chalmer's however, it is explicitly said that a being like F-Paul is a logical possibility, but that its existence is incompatible with the empirical information we have with respect to how the brain works, and in particular, with the information concerning the relation between the conscious experiences of the subject and the beliefs he forms about them. This is the strategy that I will follow in this chapter: I will argue that this version of the replacement thought experiment shows that the claim that absent qualia are empirically possible is false. In the fourth chapter of the thesis, I will present and evaluate a different strategy, offered originally by Tye in (2006), whose aim is to show that the claim that absent qualia are logically possible is false.

If we accept the *reductio* hypothesis, then the capacity of supporting conscious phenomenal mental states will disappear between some point of the sequence that exists between Paul McCartney and F-Paul. In the case of this version of the replacement

argument, we can consider two different possibilities, which in turn can be also divided in several sub-cases that can be categorized as follows:

> **Case (1)** The ability to support conscious phenomenal mental states disappears suddenly at a determinate point of sequence < J_0, J_1, J_2, , … , J_n >.
>
> **Case (2)** The ability to support conscious phenomenal mental states fades, disappearing gradually between the two extreme points of sequence < J_0, J_1, J_2, , … , J_n >.
>
> **Case (2.1)** The subject is aware of this: he notices that he loses certain conscious phenomenal states.
>
> **Case (2.1.1)** Awareness of losing phenomenal consciousness is instantiated in the part of the brain composed by artificial neurons
>
> **Case (2.1.2)** Awareness of losing phenomenal consciousness is instantiated in both parts of the brain (the replaced part and the organic part)
>
> **Case (2.1.3)** Awareness of losing phenomenal consciousness is instantiated in the part of the brain still composed by organic neurons.
>
> **Case (2.2)** The subject is not aware of this: he loses conscious phenomenal states and he does not know that he is losing them.

Case (1) The ability to support conscious phenomenal mental states disappears suddenly at a determinate point of sequence < J_0, J_1, J_2, , … , J_n >.

From this point, the argument will proceed by showing the impossibility of these cases. [Not all versions consider exactly the same cases. I've followed Chalmers's and Kirk's version] Lets consider case (1). If the *reductio* hypothesis is right, perhaps the replacement of organic neurons by artificial ones that implement the same input-output function preserves no mentality at all, but is instead a process very similar to the destruction, one by one, of the neurons of the brain. According to this initial case, there will be a point in the sequence < J_0, J_1, J_2, , … , J_n > of replacement cases in which Paul McCartney loses his phenomenal mental states abruptly when just a single biological neuron is replaced by an artificial neuron. Imagine, for instance, that the replacement process has already taken away a certain number of his original neurons and that Paul McCartney's brain is now only 73 per cent organic. Until that point, Paul has enjoyed a world of amazing phenomenal experiences, but the next time an organic neuron is removed and replaced by an artificial neuron, all these experiences disappear. According to case (1), then, the replacement of

only one neuron will mark the boundary between the complete possession of conscious phenomenal experiences, on one hand, and absolute mental death on the other.

In order to see why case (1) is implausible, it will be useful to imagine a situation in which the neurons inside Paul McCartney are destroyed one by one, and not replaced by artificial neurons. If conscious phenomenal states do not disappear suddenly when neurons are destroyed, then it is doubtful that they disappear suddenly when they are replaced by entities that preserve their same input-output behaviour. Of course, it can still be argued that these cases are different because there is certain unexpected element or property introduced by these artificial neurons, which is obviously not present in the case of neural elimination, and that it generates in some way a sudden elimination of the conscious phenomenal experiences of Paul. Before we consider this possibility, however, we need to see what happens when neurons are completely destroyed and not replaced by miniature entities that duplicate their input-output behaviour.

I want to consider two cases of neural destruction. In the first one, the neurons inside Paul McCartney's brain that are being destroyed are contiguous to each other. They may be located in any area of the brain, and the only constraint is that there must be a connection between these neurons. Imagine that the organic neurons inside the brain stem of Paul, which is the part of the brain most directly related to awareness, are removed but not replaced by artificial neurons. According to case (1), Paul McCartney would preserve his awareness when a certain number n of neurons inside his brain stem are destroyed (or extracted), but when one more neuron is taken out from his head, that is, when a number $n+1$ of neurons are destroyed or removed, he loses awareness completely. Thus, the hypothesis described by case (1) implies that Paul McCartney will transit from a state of full awareness to a state of complete unconsciousness when just a single neuron of his brain stem is removed. Note that it seems entirely plausible to say that there is a point in the replacement process in which awareness disappears completely when a single neuron is removed from Paul McCartney's brain stem, but this should be preceded by a gradual fading in his conscious abilities. Remember that the process of replacement, as it has been described, requires a gradual removal of neurons, one by one. Thus, we would expect that

Paul McCartney experienced a progressive decrease of those mental abilities that are related to awareness, and not an abrupt transition between full mental life and the complete unconsciousness. Note also that when people experience a sudden elimination of their conscious abilities, it is the activity of a significant region of the brain and not that of a single neuron what it is involved. It is clear that Paul McCartney may lose consciousness suddenly in certain circumstances: for instance, he can commit suicide by shooting himself in the head. However, this act would be the cause of a severe modification in his brain structure, and surely his losing of consciousness would not be caused by the destruction of a single neuron.

In the second case, the neurons that are removed from Paul McCartney's brain do not need to be connected among them. We may imagine that we remove organic neurons from several different regions of his brain in a random way. If this is so, it might be possible that the remaining organic neurons inside Paul McCartney's brain implement a certain adjustment process that allows him to maintain the same phenomenal conscious states in spite of the destruction of a gradually higher number of neurons. But again, case (1) still implies that the destruction of just a single neuron marks the boundary between full consciousness and a complete lack of it. Even if until a certain point the neural adjustment allows Paul McCartney to enjoy all his phenomenal conscious mental states, the cause of their elimination would be the destruction of only one neuron.

Most of the authors that offer versions of the replacement thought experiment consider this case. The crucial point is that we have sufficient empirical information to know that the destruction of a single neuron cannot have the results suggested by case (1). In (1994), Kirk evaluates this case and offers his conclusion:

> ... we know that awareness involves very many neurones working together. That being so, even total destruction of some of the neurones involved in awareness would not inevitably result in the sudden total loss of awareness [...] A fortiori, substituting miniature computers for some of the relevant neurones would not have that result.[38]

38 " Kirk (1994, p. 99)

There is, however, another possible objection that we need to consider. The only thing that we know for certain about these artificial neurons is that they implement the input-output function of an organic neuron, but we do not know anything concerning their physicochemical constitution. We simply cannot rule out the possibility that this unknown constitution affects the brain as a whole, in such a way that when a determinate number of artificial neurons are introduced in the brain, all the conscious abilities of Paul McCartney stop abruptly without fading away before this point. Or perhaps this harmful element is not located in the physiochemical constitution of the artificial neuron, but it is generated when it processes the signals it receives from other neurons. Since we simply are not sure whether the constitution or the processes inside the artificial neuron introduce this harmful element, we have no justification to claim that the replacement does not produce a sudden elimination of Paul McCartney's conscious abilities. Thus, while we can be sure that gradual neural destruction cannot bring a sudden elimination of consciousness, we cannot say the same when strange external elements, like artificial neurons, are introduced in the brain.

It is true that some physical systems that implement the input-output behaviour of an organic neuron may have harmful effects on the rest of the brain and even on the body as a whole. However, there is no reason to think that all possible implementations have these adverse effects. Suppose that artificial neurons were made out from, say, plutonium-239. Until certain point they may perform the input-output function adequately and Paul McCartney would not notice any change in his mental states. However, when critical mass is finally achieved Paul McCartney will explode, obliterating his conscious mental states in an instant together with the entire neighbourhood. Or perhaps the artificial neurons are programmed in such a way that when they reach a certain determinate number they release a harmful chemical substance that in some way damages the parts of the brain that are still organic. Before the artificial neurons reach that point the mental life of Paul McCartney might be entirely normal, but when the substance is released he immediately loses all his conscious mental states. The point I am making here, however, is that being made out from materials that have these harmful effects, or being programmed for eliminating consciousness, is not necessary for implementing the adequate input-output function of the organic neuron.

These are the reasons to think that gradual neural destruction does not bring a sudden elimination of the conscious abilities of Paul. Instead, we might expect that these abilities fade as a greater number of neurons are removed from his brain. There is good empirical support for the thesis that losing consciousness abruptly requires the modification or destruction of a large number of neurons at once, and not the removal of a single neuron. But if this happens in the case of neural destruction, there does not seem to be a good reason to think that it happens in the case of neural replacement. Even if we argue that artificial neurons may be made out from materials that have the effect of eliminating conscious abilities when a determinate number of them have been inserted in Paul McCartney's brain, there is no justification to think that all the possible materials from which artificial neurons are made have this harmful effect. This, for now, exhausts case (1).

Case (2): The ability to support conscious phenomenal mental states fades, disappearing gradually between the two extreme points of sequence $< J_0, J_1, J_2, , \ldots , J_n >$.

So far, we have seen that there is no good reason to think that conscious mental states disappear when only one neuron is destroyed, so we have no reason to think that these states disappear when a single neuron is replaced by a different physical entity that implements its same input-output behaviour. This leaves us with case (2), where the conscious phenomenal states of Paul fade away as the replacement process is being performed. As we have mentioned before, this case can be divided into another two exhaustive sub cases: in case (2.1), Paul McCartney is aware of losing his conscious phenomenal states, while in case (2.2), he is not aware of losing anything. Thus, our new *reductio* assumption can be expressed as follows:

Case (2.1): Paul McCartney is losing his conscious phenomenal states, and he is aware of this loss.

Part of the protocol of the replacement operation consists in that the scientists, with the help of a microphone installed inside the machine, ask Paul questions concerning his physical

state and his current perceptions: “Are you feeling all right?”, “Do you hear your own voice?”, “Can you see the little microphone in front of you?”. The scientists do this in order to check if the replacement process is performed adequately. If Paul noticed something odd with his visual or auditory perceptions, he would be able to inform this to the group of scientists. The scientists could then check if there is something wrong with the procedure. Perhaps some artificial neurons do not work as they are supposed to, or maybe the scientists inserted by mistake some artificial neurons in the piriform cortex that were originally conceived as replacements of the organic neurons inside the visual cortex. (Notice that such procedures are frequent in neurosurgery: in some cases, the patient remains conscious during the operation, in order to know whether there is something wrong).

But imagine that, at some point, the patient notices that his current phenomenal experiences gradually disappear. It might not be clear, however, the precise way in which Paul´s phenomenal experiences fade. A first option is to imagine that scientists decided to start the replacement process by extracting the organic neurons from some region of the brain associated to a certain perceptual ability. For instance, assume that the scientists started by replacing the neurons located inside Paul´s primary auditory cortex, which is the region associated with the perception of sounds. If the reduction hypothesis were true, then Paul would notice that, as the replacement process goes on, his auditory perceptions become gradually weaker. Another option would be to imagine that the machine replaces the neurons in a random way. In any case, Paul´s phenomenal experiences would gradually become weaker as the replacement process goes on. Imagine now that Paul is singing Eleanor Rigby inside the replacement machine, and in the middle of the song he notices that his voice, together with the sound of his bass, becomes fainter. At that moment, one of the scientists asks him: “Is everything all right, Paul?” Evidently, Paul wants to tell him that there is something wrong and that the replacement process resulted in a modification in his auditory perceptions. He might have an attack of panic, screaming ad telling the scientists that they did something terribly wrong and that he is now unable to hear his own voice, the voices of the scientists or the sound of his bass.

Another option it is worth to consider is the possibility that some of the cognitive capacities of Paul fade during the replacement process. Consider, for instance, the case of memory. While Paul is sitting inside the replacement machine, he may become ignorant of what he is doing there, surrounded by several unknown men in white robes. Depending perhaps on the number of organic neurons that have already been replaced, he might also forget some details concerning important events in his past: he might have lost the memory that he was part of a band with some friends several years ago, that he was married once to a woman named Linda, or even that his own name is Paul. Like the previous case concerning his auditory perceptions, he might stand up, looking around confusingly and asking to himself "Who am I?", "What is this strange place?", or "Why are these men in white robes speaking to me?".

The reaction exhibited by Paul after noticing that there is something wrong with his auditory perceptions would probably be telling the scientists that he thinks they are doing something wrong and that the replacement process is not working as planned. Also, if the memories of Paul were gradually lost – for instance, if as a result of the replacement process he forgot that he is inside of the replacement machine, that he is surrounded by scientists in white robes, or even that he forgot his own name – he might cry for help, asking things like "What I am doing here?", "Who are you, people?", or "Who am I?". In any case, the assumption is that his mental capacities are fading would be reflected in his behaviour.

Note again that Paul´s behaviour after the replacement of the organic neurons of this region of his brain is going to be identical to the one he would have exhibited in case that the replacement had not been performed. The reason is that the motor neurons in his brain are going to send exactly the same signals to the muscles in his body that the organic neurons would have sent. In the same way, his speech, which also depends on the movements of his body, would be exactly the same. Thus, Paul would not be able to tell the scientists that there is something wrong with the replacement process.

Of course, the fact that the members of the sequence will exhibit the same behaviour as Paul after the replacement does not imply, by any means, that his mental capacities, and

in particular the capacity of experiencing conscious phenomenal states, are preserved after the process of replacement. The closer we get to the other extreme of the sequence, that is, to the point where F-Paul is situated, the more difficult is to accept that these mental capacities are the same. While Paul is a person who enjoys all kind of phenomenal mental states and is able to experience pain, sexual arousal, or the colour of a blue parrot, a being like F-Paul – which instead of a brain has something whose only resemblance to Paul´s brain is that it shares its functional organization at a neural level – is no more than a zombie (or better, a functional zombie). This being looks like Paul, moves like him, says the same things and it may even be as fun as he is, but lacks any sort of mental life.

In (1992), Searle examines the outcome of a very similar scenario in which a patient loses his visual abilities as a result of the replacement of his organic neurons by entities that implement the same input-output behaviour but that are made out from silicon. This is how Searle evaluates this scenario:

> … as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behaviour. You find, to your total amazement, that you are indeed losing control of your external behaviour. You find, for example, that when doctors test your vision, you heard them say, "We are holding up a red object in front of you; please tell us what you see." You want to cry out, "I can't see anything I'm going totally blind." But you hear your voice saying in a way that is completely out of control, "I see a red object in front of me.[39]

Is this an acceptable picture? Note also that, under such circumstances described in the last paragraph, Paul would surely start to experience an emotion of fear, even close to extreme horror as he realizes that he is unable to control his answers to the questions the scientists ask concerning his auditory perceptions. Furthermore, his entry to the scenario would have been entirely involuntary, because he didn't perceive the sound and he obviously does not want to appear out of time, so he would realize that he is not able to control the movements of his body. The question is, how can this awareness of losing his auditory abilities and the feelings of terror that such awareness may trigger be instantiated

39 " Searle (1992, pp. 66-67)

by Paul's brain? It seems that we have here three different options. Either (2.1.1) awareness is instantiated by the artificial neurons that have replaced the organic ones; (2.1.2) they are instantiated both by artificial neurons and organic neurons, or (2.1.3) they are entirely instantiated by the organic neurons that are still inside Paul´s brain.

Options (2.1.1) and (2.1.2) can be easily ruled out. The reason is that this would imply that artificial neurons are, after all, capable of supporting conscious mental states. The only option left is, thus, (2.1.3): awareness and related emotions are instantiated by the organic neurons that have still not been replaced.

But this is not possible for the following reasons. One of the assumptions of the replacement picture is that the part of Paul´s brain that remains composed of organic neurons is in the same state it would have been in case that the replacement had not been performed. In order for new beliefs to be formed, in order to be aware of losing the ability of perceiving auditory stimuli and to experience feelings and emotions related to it, Paul McCartney's brain needs to be in a certain state. But by the assumptions of the thought experiment, his brain is in the same state that it would have been if the rest of his neurons had not been replaced. It is for this reason that Paul McCartney cannot experience these new mental states. There does not seem to be any place for these states to be instantiated. Chalmers's evaluation of the outcome proposed by Searle is clear in this point:

> There is simply no room in the system for any new beliefs to be formed. Unless one is a dualist of a very strong variety, this sort of difference in belief must be reflected in the functioning of a system—*perhaps* not in behavior, but at least in some process. But this system is identical to the original system (me) at a fine grain. There is simply no room for new beliefs such as "I can't see anything," new desires such as the desire to cry out, and other new cognitive states such as amazement. Nothing in the physical system can correspond to that amazement. There is no room for it in the neurons, which after all are identical to a subset of the neurons supporting the usual beliefs […] Failing a remarkable, magical interaction effect between neurons and silicon—and one that does not manifest itself anywhere in processing, as organization is preserved throughout—such new beliefs will not arise. [40]

40 " Chalmers, (1996a, p. 258)

Thus, Paul McCartney cannot experience these new mental states, and therefore, we can rule out case (2.1). At this point, however, one may be tempted to argue as follows: "You argument is fallacious because you assume, without proof, that the electrochemical signals transmitted among neurons exhaust mentality. This is precisely what the functionalist needs to say in order for the replacement argument to be successful, but you have not ruled out the possibility that some other physicochemical event in the brain is supporting Paul McCartney's awareness of losing his ability of hearing the sound, as well as the emotions of terror that he may experience when he realizes that he is unable to control his movements. Such an event may even be instantiated *inside* the neural body, and not related to the signals that neurons are interchanging among themselves. You have not ruled out this possibility, and thus, you are simply begging the question when you assume that the electrochemical signals exhaust mentality."

It is true that the argument, as it has been presented here, does not rule out the possibility of such event, but we do not need to do it at this stage of the argument. However, we do not need to commit to the idea that the electrochemical signals exhaust mentality. The reason is that the part of the brain that has not been affected by the replacement, that is, the set of organic neurons that is still left, is exactly in the same state it would have been in case that the replacement was not performed. So any event, including all the physicochemical events that may happen inside the organic neurons, are essentially the same. There is no difference at the functional level, but it should be clear that there is no difference at the biochemical level either. The part of the brain that is still organic has not been changed, and this is the reason it cannot instantiate Paul McCartney's awareness or his emotion of fear.

For these reasons, I think we can rule out case (2.1.3), which in turn exhaust case (2.1). Lets examine now the last case proposed at the beginning of this section.

(2.2) The conscious experiences of Paul fade along the replacement sequence, but he is unaware of this.

Consider again the scenario illustrated by Paul´s operation, and imagine that a large number of the organic neurons located in his primary auditory cortex have been replaced by artificial neurons. Assuming the truth of the hypothesis that his conscious experiences fade along the replacement sequence, this percentage is probably sufficient for a notorious decrease in the intensity of his auditory experiences. Note that the human capacity of hearing high-frequency sounds diminishes with age (this phenomenon is known as presbycucis). The causes of this phenomenon are not entirely clear, but it might be related, precisely, to the progressive loss of neurons. Imagine now that one of the scientists in charge of monitoring the replacement process evaluates Paul´s reaction to a high-pitched sound over 12kH, and assume that the replacement process is at a point where Paul can still hear the high-pitched sound, but since his auditory experiences are fading, the sound he in fact hears is comparatively weaker that the one he would hear in normal circumstances. Before the sound is produced, the scientist asks Paul the following: "Paul, the sound you are going to hear is very loud. Please tell me if you can hear it clearly." After the sound is produced, and no matter the intensity of his auditory experience, Paul would say the scientists that he heard the sound with clarity. The reason, again, is that the replacement process does not modify his behavioural dispositions. But since the sound he hears is extremely weak, this answer would be mistaken. Of course, if Paul noticed that he gave a wrong answer to the question of the scientist, he would notice that there is something wrong with his auditory experiences. But the examination of the last case showed that this is not possible. Paul cannot notice that his auditory experiences disappeared or that they are comparatively weaker compared with the experiences he had before the replacement process started.

Cuda argues against the claim that the experiences of the subject are not eliminated by the replacement process in this way: if it makes sense to say that Paul can be so mistaken about his current experiences, then it makes sense to say that we can also be so mistaken. Consider the following example: the Mosquito alarm is a device that emits a powerful high-frequency sound that can be heard only by people under certain age (mostly people under 25 years old). The alarm is advertised as a method to stop loitering by youths: the sound it produces is so powerful and annoying that, according to its designers, deters

people from congregating in the areas where that sound can be heard. Imagine now that a normal teenager whose brain is not damaged hears the alarm. When he hears that sound, he forms the belief that there is a Mosquito alarm in the vicinity and that the sound it produces is extremely annoying. He also forms a strong desire of not hearing that sound, and consequently, he runs away from that place. Both the teenager´s behaviour and the cognitive states he forms are consistent with him hearing the annoying sound of the Mosquito alarm. Now, the question is whether the teenager can behave and form beliefs and desires in a way that is consistent with hearing the annoying sound emitted by the alarm, but without hearing it at all. Remember: the teenager´s brain is completely normal, and he is a completely rational person. According to Cuda, it does not make sense to say that a rational person can be so mistaken about his own experiences. He exemplifies this case with a functional isomorph able to form beliefs and desires concerning red qualia without being able to experience them:

> If it makes sense to think that [a functional isomorph] could be mistaken in such a way, then it makes sense to think that we could be mistaken in such a way also. Hence we would have no reason to think that things made of neurons (i.e., ourselves) have red qualia. But clearly it makes no sense to think that we act like we have red qualia, believe that we have them, etc., but that we are all mistaken and really never have any red qualia. Therefore, it makes no sense to think that [a functional isomorph] could be mistaken in this way either.[41]

However, it can still be argued that Paul´s case is different from the one just described. Why not to say that when the replacement process is performed Paul loses not only the capacity of hearing this sound, but also the capacity of forming beliefs and other cognitive states about it? Perhaps one of the effects of the replacement process is that Paul simply becomes more and more stupid. When artificial neurons take the place of organic ones he gradually becomes incapable of understanding his own behaviour and the perceptual experiences he still might have. In particular, he might be unable to form beliefs concerning the capacity of hearing sounds over 12kHz. In such a case, he could not form the belief that he answered affirmatively to the question "Paul, did you hear the signal?"

41 " Cuda, (1985, p. 117)

The argument proposed by Chalmers shows that the situation envisaged in the last paragraph is extremely implausible. In (1996 and 2010), Chalmers argues in favour of what he calls the principles of Structural Coherence. According to these principles, the phenomenal aspect of the mind and the psychological aspect are systematically related. In particular, Chalmers proposes what he calls the *Reliability Principle*: our second order-judgements concerning our conscious experiences are by and large correct. For instance, when a normal subject that is paying attention to his experiences and that does not suffer from a neurophysiological illness judges that he had a visual sensation, he usually has a visual sensation. Also, when this subject forms the belief that he has a strong, painful experience in his right toe, he usually has a painful experience in his right toe. Conversely, the *Detectability Principle* suggests that when a being has a conscious experience, in general he has the capacity of forming a second-order judgement about this experience.

> ... our second-order judgments about consciousness are by and large correct. We can call this the *reliability* principle. When I judge that I am having an auditory sensation, I am usually having an auditory sensation. When I think I have just experienced a pain, I have usually just experienced a pain. There is also a converse principle, which we might call the *detectability* principle: where there is an experience, we generally have the capacity to form a second-order judgment about it.[42]

These principles are not conceived by Chalmers as logically or conceptually necessary: our second-order judgements can be sometimes mistaken. For instance, the subject can be inattentive to his conscious phenomenal experiences: a distracted subject may confuse an orange experience with a yellow one. He may suffer from congenital analgesia, a rare condition that prevents a patient from feeling physical pain. In this case, some subjects can judge, mistakenly, that they are in pain. But this may be explained by suggesting that the individual is unaware of his condition.

That Paul does not notice that his auditory experiences fade along the replacement and, at the same time, he does not form the belief that there is something wrong about these experiences clashes with the principles proposed by Chalmers. When Paul reflects on his

42 " Chalmers (1996, pp. 218-219)

experiences, he would form the belief that he heard a clear sound, in spite of having only a very weak auditory experience. If the absent qualia hypothesis were right, the extreme point of the replacement sequence (that is, F-Paul) might perhaps have lost not only the ability of having any sort of auditory experiences, but also the capacity of forming beliefs about his own behaviour. But remember that in the case that we are evaluating, when only a certain percentage of the neurons inside his primary auditory cortex have been replaced by artificial neurons, Paul is still not a complete zombie. Although his capacity of hearing sounds has been seriously diminished, he is still capable of hearing the sound produced by the scientists, and moreover, he is still capable of reflecting about his own behaviour. He would believe that he understood the instructions the scientists gave him, and that he heard a clear sound instead a weak one. Also, he would believe that and that he answered affirmatively to the question of the scientists. But if he was unable to hear the high-pitched sound, there is no way of explaining how these beliefs have been generated.

> Here we have a being whose rational processes are functioning and who is in fact *conscious,* but who is utterly wrong about his own conscious experiences. Perhaps in the extreme case, when all is dark inside, it might be reasonable to suppose that a system could be so misguided in its claims and judgments—after all, in a sense there is nobody in there to be wrong. But in the intermediate case, this is much less plausible. In every case with which we are familiar, conscious beings are generally capable of forming accurate judgments about their experience, in the absence of distraction and irrationality. For a sentient, rational being that is suffering from no functional pathology to be so systematically out of touch with its experiences would imply a strong dissociation between consciousness and cognition. We have little reason to believe that consciousness is such an ill-behaved phenomenon, and good reason to believe otherwise.[43]

But why not say that the judgements made by Paul about his conscious experiences are extremely wrong? Consider, for instance, the case of patients that suffer from blindness denial (also called the Anton-Babinsky syndrome). This syndrome is a particular case of a disease known with the name of "anosognosia", which is the denial of illness that, in some cases, can be seen in patients exhibiting brain injuries. In the case of the Anton-Babinsky syndrome, the eyes of the patients are perfectly capable of responding to light. However,

43 " Chalmers (1996, p. 257)

injuries to the visual cortex prevent these patients to visually discriminate objects, shapes and colours. In spite of this, patients usually do not accept having any visual difficulties. It can be argued that Paul´s case is very similar to the one of a patient suffering from blindness denial. Perhaps it is true that Paul´s experiences are diminishing along the replacement process, but he simply cannot realize that there is something wrong with these experiences.

Now, a possible explanation for the Anton-Babisnky syndrome is that the damage to the visual cortex, apart from causing blindness in the patient, also prevents an adequate communication with the parts of the brain that control the verbal behaviour of the subject. By hypothesis, this could not happen to Paul. To begin with, the explanation of the Anton-Babinsky syndrome involves the fact that the communication between a damaged visual cortex and the parts of the brain in charge of the verbal behaviour of the subject is impaired. But the only way in which this is possible is through a difference in the way that the neural replacements affected the regions of the brain that control this verbal behaviour. By hypothesis of the replacement process, this is not possible: at the synaptic level, the artificial neurons affect these areas in exactly the same way as organic neurons. For these reasons, we can conclude that the case of Paul is different from the case of a patient that exhibits blindness denial.

If we accept that awareness is a form of conscious experience, then it is possible to claim that Paul's awareness of his conscious phenomenal experiences fades in parallel with respect to the fading of his object-level conscious experiences. But note that an individual can be in a conscious state despite the removal of all the neurons in their primary visual cortex. Thus, the replacement of the neurons in the visual cortex would not modify the global conscious state of the subject.

Lets recapitulate the argumentative strategy of this section. To assume that the conscious experiences of the subject disappear along the sequence of replacement cases generates two different possibilities: the first is that the ability of generating conscious phenomenal states disappears at a determinate point of the sequence. The second is that this

ability fades between the two extreme points of the sequence. The first possibility was dismissed on the basis that we have good empirical reasons to think that consciousness does not suddenly disappear when a single neuron is eliminated, so it is also extremely implausible that the replacement of a single organic neuron by an artificial one brings with it the elimination of the conscious experiences of the subject. The second possibility was divided into two subcases: in the first subcase, the subject is aware that his conscious experiences fade along the sequence of cases. In the second subcase, the subject is not aware that his conscious experiences fade. The subject cannot be aware that his experiences are fading because this awareness needs to be instantiated in some part of his brain. It cannot be instantiated in the part that remains organic: by hypothesis, this part is receiving the same signals from the part of the brain composed by artificial neurons, so it is in the same state it would have been in case that the replacement had not been performed. But this awareness cannot be instantiated by the parts of the brain composed by artificial neurons, because this would mean that they are, after all, capable of generating conscious states. This brings us to the final subcase: the conscious experiences of the subject fade along the replacement sequence, but he is unaware of this. The reason we have presented against this possibility is that it would imply that the beliefs that the subject forms concerning his conscious experiences are systematically wrong. However, the principles suggested by Chalmers show that this is extremely implausible: according to the Reliability Principle, the second-order judgements of the subject are usually correct. But in the case illustrated by Paul, he would form the belief that he heard a clear sound, while in fact the sound heard by him is very weak. If we accept the claim that the experiences of the subject cannot be dissociated in this way from the beliefs he forms about them, then we can reject the claim that the subject ignores that his experiences fade along the replacement sequence. The cases evaluated exhaust all the possible ways in which the experiences of the subject disappear along the replacement sequence. Thus, the absent qualia hypothesis is, at least, naturally false: it is not the case that a system that duplicates the functional organization of the brain of a conscious being at a neural level lacks the capacity of generating conscious phenomenal experiences. But if this is true, thesis (NSS) is also true: then there is at least one functional organization – namely, the functional organization of the brain at a neural level – such that the property of a system P of generating conscious phenomenal

experiences naturally strongly supervenes on the property of implementing this functional organization.

Why this version of the replacement thought experiment does not show, instead, that the preservation of the functional organization of the subject´s brain preserves the property of generating conscious phenomenal experiences with logical necessity? Note, first, that the argument presented in this chapter assumed the truth of certain empirical facts about the brain in order to dismiss the idea that phenomenal consciousness was not preserved along the replacement process. In order to argue against the claim that the conscious experiences of Paul do not disappear suddenly when a single neuron is replaced, we considered empirical information concerning the fact that the generation of conscious experiences involves a large number of neurons. Note that we cannot rule out *a priori* the possibility that a single neuron, in some strange and undiscovered way, played a crucial role in the generation of consciousness, in such a way that its removal generates an immediate and complete destruction of the conscious experiences of the subject. Of course, the empirical information we possess rules out the presence of such a neuron. Second, the assumption that Paul could not form the judgement that there is something wrong with his auditory experiences clashes with the principles of Structural Coherence proposed by Chalmers: In particular, this clashes with the Reliability Principle, according to which his second-order judgements about his conscious experiences are usually correct. Paul would believe that he heard a clear sound, and that he gave an affirmative answer to the question "Did you hear the sound clearly?" But if his auditory experiences are fading, the sound he heard was very weak, and thus, that he formed the belief that he heard a clear sound clashes with the Reliability Principle. But neither the Reliability Principle nor the Detectability Principle are logically necessary. As it has been mentioned, the second-order judgements of a subject can sometimes be erroneous.

> The principles I have outlined are not absolute. Our second-order judgments can sometimes go wrong, providing exceptions to the reliability principle. [...] In the reverse direction, it is arguable that experiences can be unnoticeable if they occur while one is asleep, for example, or if they flicker by too fast for one to attend to them. But all the same, these principles at least encapsulate significant regularities. In a typical case, a second-order judgment will usually be correct, and an experience will

usually be noticeable. These regularities are not exceptionless laws, but they hold far too often to be a mere coincidence. Something systematic is going on.[44]

In the fourth chapter of this thesis, I will present a different version of the replacement thought experiment that is explicitly conceived as showing that the absent qualia hypothesis leads to a logical or conceptual contradiction. If this is true, then the hypothesis of absent qualia would be logically false, and not only empirically false.

Section 4

Objections to the neural version

Remember that one of the worries concerning the *regions-of-the-brain* replacement scenario was that it is difficult to conceive a device that shares all the causal relationships that the visual cortex has with respect to the rest of the brain but whose physiochemical composition is radically different from it. Nonetheless, the scenario can be adequately described by considering a device that duplicates only the input-output neural pattern of the visual cortex and not all its properties. But how feasible is this duplication? In order to give an answer to this question, it will be useful to remember, briefly, how a neuron produces and transmits signals to the neighbouring neurons in the brain. A possible source of scepticism concerning the replacement argument in general may be due the perceived omission of scientific data concerning the way a real neuron works by the proponents of replacement arguments.

The authors postulate the existence of physical entities capable of duplicating the input-output behaviour of a neuron, but in general they simply do not give an account of how can this be accomplished, and this leaves the impression that the description of the replacement scenario leaves some essential points unexplained. One of the objectives of this section is to show, in a very general way, how a neuron functions and how it produces the electrochemical signals used for communicating with other parts of the brain and the body. Of course, it is far from the objectives of this thesis to explain how a neuron works

44 " Chalmers (1996, p. 219)

with the level of precision achieved by contemporary neuroscience. However, it is possible to have an adequate general picture of this process and to imagine how it can be, perhaps, implemented by a fictional (although potentially real) non-biological physical system. This is how a contemporary, scholarly text in neurophysiology defines, in very general terms, what a neuron does:

> Neurons are remarkable among the cells of the body in their ability to propagate signals rapidly over large distances. They do this by generating characteristic electrical pulses called action potentials or, more simply, spikes that can travel down nerve fibers. Neurons represent and transmit information by firing sequences of spikes in various temporal patterns. [45]

All cells in the body maintain a certain voltage potential, but in the case of neurons, this plays a basic role in the way that they generate the signals that communicate them with other neurons and cells of the body. When a neuron is in its resting state, there is an average of -70 millivolts of voltage difference between the interior part of the neuron and the external medium. This voltage difference is regulated by ion channels, which are protein-like structures located in the neural membrane that function like valves, and are sensitive to the electrical properties of the electrically charged particles that float inside and outside the neuron. For instance, sodium ions (Na+) are in higher concentrations outside a neuron than inside it, and conversely, the concentration of potassium ions (K+) is higher inside the neuron and lower in the outside.

It is precisely the interchange of these electrically charged particles through ion channels that create this voltage difference between the inside and the outside of the neuron. Neurons are stimulated at several points of stimulation at the dendrites. When this stimulation meets a particular threshold, the neurone generates electrical pulses known as "action potentials" (sometimes also called "spikes"), which are signals that travel across the axon and that stimulate other neurons in the brain or muscle cells in the whole body. When electrical current in the form of positively charged ions flows out of the neuron, or when negatively charged ions flow into the neuron, the electric potential of the neuron's

45 " Dayan & Abbott (2001, p. 3)

membrane becomes more negative. Conversely, when negatively charged ions flow out of the neuron, or when positively charged ions flow into the neuron, the electrical potential of the neuron becomes more positive. These processes are known, respectively, as hyperpolarization and depolarization. When a neuron is depolarized above a certain threshold level, the neuron generates an action potential. Action potentials are temporal fluctuations in electrical potential across the membrane of the neuron (of about 100 mV and whose duration is one milisecond) and can be transmitted over large distances through the axon. When the action potential reaches the synaptic connections at the end of the axon, this produces a new opening of ion channels that lead to the release of a signal through the electrical or the chemical synapses of the neuron.

This is, in a very general way, how a neuron transmits electrochemical signals to other neurons and cells in the brain and the body. With this in mind, we can now try to give a sense of the idea of building artificial devices able to implement the input-output causal behaviour of a neuron. As we have seen, the input-output causal behaviour of a neuron is a very complex process that involves the generation and transmission not only signals of an electrical character – the action potential. It also interchanges sodium and potassium ions with the medium that surrounds the neuron, together with neurotransmitters at the synapses. When the replacement scenario is described as including artificial neurons that implement this same input-output behaviour, we need to suppose that all that the neuron produces is also produced by the artificial neuron, and that it is adequately transmitted to the neighbouring neurons at the right time.

The details of this process might suggest that creating artificial neurons from any imaginable materials is extremely unlikely. Artificial neurons composed entirely of Putnam's Swiss cheese, for instance, are obviously ruled out. Again, we are at risk of saying that the only replacements capable of duplicating the input-output behaviour of a neuron – their synaptic behaviour – are gadgets with an identical physicochemical composition of organic neurons. And we have the same problem as before: even if consciousness is preserved after the replacement, it is not possible to rule out the possibility that the chemical composition of the replacements is essential for generating consciousness.

The second problem is similar to the one we have discussed in the last section. How different can be the physiochemical constitution of an artificial neuron from an organic one? If we stipulate that the replacements need to duplicate all the causal relations that organic neurons have with respect to the rest of the brain, it might be that only gadgets whose physiochemical constitution is very similar to organic neurons can perform the task adequately. Again, we can mention properties like weight and temperature. Organic neurons and artificial ones that include components made from stainless steel will interact in a different way when exposed to magnetic fields, for instance.

If the replacements are defined as duplicating all the causal relations that organic neurons have with respect to the rest of the brain, the strategy may succeed in showing that the resulting physical system – the one composed entirely by artificial neurons – is able to generate phenomenal consciousness. Nonetheless, we have argued that a device that duplicates all these causal relations without having, at the same time, a very similar physiochemical composition of organic neurons is extremely difficult, if not impossible. Moreover, remember that organic neurons interact with their surrounding medium in several ways. As we have mentioned, neurons are surrounded by a "bath" of electrically charged ions. These ions flow through the membrane of the cell in a process known as the "sodium-potassium pump", which contributes to the generation of the action potential produced by the neuron.

In spite of these observations, notice that we may describe the replacement scenario by stipulating that the only causal interaction that needs to be considered is the one related to synaptic transmission. Organic neurons are able to communicate among themselves through chemical and electrical synapses, and this process is regulated by mechanisms that are located in the interior of the neuron. However, a neural replacement does not need to generate the adequate signals in the same way.

Remember the case of organs like artificial hearts or hips: they perform the right function without sharing all the causal properties of the original organs. Similarly, we can

imagine physical devices able to perform the essential function of organic neurons, without sharing, at the same time, all their causal relationships with the rest of the brain. As we have seen earlier, the sodium-potassium pump is essential for generating the action potential in organic neurons. But the process under which neural artificial replacements generate the adequate signals does not need to be identical. Artificial neurons may be equipped with tiny devices that generate electrical signals according to the instructions of a program, and may also possess a miniature factory that is in charge of producing chemical neurotransmitters. Perhaps these artificial neurons include devices that allow them to interact with the surrounding neural medium by interchanging charged ions, in such a way that the whole process does not inadvertently affect other regions of the brain. In this way, we can argue as before: it is possible to build artificial neural replacements that duplicate the synaptic behaviour of organic neurons without assuming that these replacements share all the physical properties of organic neurons.

Like the case discussed in the second section of this chapter concerning a gadget that replaces the visual cortex of the subject, there does not seem to be any *a priori* difficulty for building artificial neurons that duplicate the signals transmitted by organic neurons. But the problem, again, is that this approach can lead to a *naive* understanding of the replacement strategy. Note that the replacement of all the organic neurons inside the brain by artificial ones will generate a system that is functionally isomorphic to the brain at the neural level. But it is clear that not only the functional organization of the brain was preserved by the replacement process. The electrochemical character of the signals interchanged by organic neurons was preserved as well. The problem, of course, is that in this case it is not possible to assure that the generation of conscious phenomenal experiences depended on the functional organization of the system and not on the particular physical character of the transmitted signals. I do not mean, however, that the authors that have offered versions of the replacement thought experiment understand it in this way. In this section I will show how Chalmers and Cuda explain the replacement process without appealing to the preservation of the electrochemical character of the signals transmitted by organic neurons.

The replacement scenario can be described without stipulating that the signals interchanged by artificial neurons have the same physiochemical composition of the signals transmitted by organic neurons. In fact, the versions of the replacement scenario described by Cuda and Chalmers envisage a situation in which the electrochemical signals interchanged by neurons are replaced by other signals. Note that in the initial stages of the replacement process – for instance, when only one neuron has been replaced by an artificial neuron – these electrochemical signals need to be preserved simply because there is no other way in which the replacement device can causally affect the rest of the organic neurons. Consider, for instance, the second member of the sequence, that is, a being that is almost identically to Paul except that it has only one organic neuron replaced by an artificial neuron. After the artificial neuron takes the place of the organic one, it restores all the original connections and transmits the adequate signals to its neighbouring neurons. At this stage of the replacement process, these signals need to have the same physical constitution of the ones interchanged by biological neurons, since there is no other way for communicating with the rest of the organic brain. However, in later stages of the replacement process, when two neighbouring neurons are replaced by corresponding artificial neurons, it may not be necessary to preserve these electrochemical signals. In fact, when all the neurons in the brain are replaced by artificial neurons it is not necessary to preserve any of these electrochemical signals in order for the system to duplicate the functional organization of the brain. If the artificial neurons are tiny computers, they might communicate by receiving and transmitting luminous impulses, or perhaps by a mutual manipulation of cogs and levers in the surface of the artificial neuron, or by any conceivable process. Cuda describes the final result of the replacement process as follows:

> Finally, after a trillion or so operations, there is nothing left of the original matter of [the] brain. At this point, most of the homunculi don't do anything with neurons anymore and have put away their neuron manipulators. Instead, they operate only between themselves, calling out what would have been the state of the neuron that they replaced. [46]

46 " Cuda (1985, p. 112)

Note that when the replacement process is described in Cuda's way, not only the physicochemical composition of the replacements – and consequently, of the whole system – has been dispensed with. The signals formerly interchanged by organic neurons, and that were also transmitted by the homunculi during the initial stages of the replacement process, also have disappeared. What we have now is a collection of homunculi that "call out" the states of the neuron that have been replaced. Perhaps these homunculi developed a process by which each of the possible physical states of the neuron receive a name, and when one of these homunculi have been received the information of the "states" of all its neighbours, it would process this information in such a way that the resulting message transmitted by it is precisely the name of the corresponding neural state. Chalmers proposes a similar strategy:

> … once both [neurons] are replaced we can dispense with the awkward transducers and effectors that mediate the connection between the two chips. We can replace these by any kind of connection we like, as long as it is sensitive to the internal state of the first chip and affects the internal state of the second chip appropriately (there may be a connection in each direction, of course). Here we ensure that the connection is a copy of the corresponding connection in Robot; perhaps this will be an electronic signal of some kind.[47]

Conclusions

The first version of the replacement argument that was evaluated in this chapter was the *regions-of-the-brain* version. A gadget that preserves the same causal relations that an organic region of the brain associated to certain mental function has with the rest of the brain (in this case, the visual cortex) will affect the rest of the brain in exactly the same way as this organic region. In particular, the regions responsible of the linguistic behaviour of the subject will be affected in the same way by this gadget. Thus, the linguistic behaviour of the subject will be identical to the behaviour he would exhibit in case that the original visual cortex had not been replaced. But it is absurd to think that the behaviour of the subject was preserved and, at the same time, the replacement eliminated the conscious phenomenal experiences formerly generated by the replaced region. While it is true that

47 " Chalmers (1996, p. 254)

Zuboff noticed something problematic in the fact that a functional isomorph at the level considered can behave in a way that is consistent with the presence of conscious phenomenal experiences but without having them at all, this does not seem sufficient to show that the absent qualia hypothesis is false.

In the second section of this chapter, I claimed that if we define the gadget that replaces a region of the brain as duplicating all the causal relationships that this region had with the rest of the brain, then it is possible that only a biologically identical gadget can duplicate these relationships without introducing any unintended effect on the rest of the brain. But if the objective of the replacement thought experiment is to support the supervenience thesis, it is doubtful that the thought experiment achieves this goal. The reason is that we cannot rule out that a property related to the physiochemical composition of the gadget is responsible for generating the conscious phenomenal experiences of the subject. I argued that the replacement scenario could be defined without stipulating that these artificial replacements share all the causal relationships of the organic brain regions. It can be stipulated that these devices duplicate only these neural properties of these organic regions.

The regions-of-the-brain version of the replacement thought experiment might be objected to on the basis that the damage of a particular region of the brain does not necessarily bring with it the elimination of the mental function associated to this particular region. The phenomenon known as the plasticity of the brain shows that, in some cases, the location of the neural activity linked to some mental function can be relocated in other regions of the brain. In particular, it can be argued that Paul preserves his visual experiences after the replacement not in virtue of the replacement of his visual cortex by a gadget that preserves the functional organization of his brain, but simply because the neural processes related to vision were relocated to a different place in his brain. Although this is not conceived as a definitive objection against the *regions-of-the-brain* version of the replacement thought experiment, it motivates the consideration of a different replacement scenario that contemplates a more detailed level of functional organization.

The second version of the replacement thought experiment presented in this chapter was the neural version. We have seen that, if it is assumed that the absent qualia hypothesis is true, the way in which the conscious experiences of the subject disappear along the replacement sequence $< J_0, J_1, J_2, , \ldots , J_n >$ can be divided into two main cases. The first was that the conscious phenomenal states of the subject are completely eliminated at a determinate point of the sequence. However, there are good empirical reasons that show that consciousness is not completely eliminated when a single neuron is destroyed, so it is extremely unlikely that the replacement of a single neuron had this effect. The second case was divided into two subcases: either the individual is aware that his experiences fade, or he does not notice it. We argued that this awareness cannot be instantiated in the organic part of the brain of the subject, nor in the part of the brain already composed by artificial neurons. Finally, the last possibility was that the conscious experiences of the subject fade, but he is not aware of this. In this case, his judgements concerning these experiences would be systematically wrong. But if we accept that the experiences of the subject cannot be disconnected in this way from the beliefs and judgements he forms about them, then we can reject the claim that the subject ignores that his experiences fade along the replacement sequence.

My final claim in this chapter is that the neural version of the replacement thought experiment adequately supports thesis (NSS). The objections I presented against this version of the replacement thought experiment are the following: perhaps the judgements that the subject forms with respect to his phenomenal experiences during the replacement are extremely wrong, similar to the judgements made by patients affected by blindness denial. However, any difference in the verbal behaviour of the subject would imply a difference in the way that the neural replacements affect the parts of the brain in charge of this verbal behaviour. By hypothesis of the replacement scenario, this is not possible. Finally, in the last section of this chapter, I briefly described the process in which an organic neuron generates and transmits electrochemical signals to other parts of the brain. I argued, first, that an artificial neuron does not need to generate these signals in the same way as an organic neuron. Second, I have shown that the electrochemical signals

transmitted by organic neurons do not need to be preserved in order for generating a system that preserves the functional organization of the brain of the subject.

Chapter 3

The replacement thought experiment and the problem of universal instantiation

Introduction

As the replacement process has been described in the last chapter of this thesis, the devices that replace whole regions of the brain or organic neurons may be constructed from a number of different materials. Of course, we have seen that there are certain constraints. For instance, the materials cannot introduce unintended effects on the rest of the brain that might impair the way in which the whole system works. Also, they need to be more or less resistant, in such a way that the gadgets may transmit signals with the adequate speed. However, as I also mentioned in the previous chapter, there is a problem related to the nature of the inputs and outputs that are processed and produced by the system. As the replacement process has been described here, the input and output signals interchanged by artificial neurons preserve the same physical nature of the signals interchanged by organic neurons. As we have seen, these inputs and outputs have an electrochemical character. Neurons generate electrochemical impulses that are generated by the flow of electrically charged ions through the membrane and the axon of the neuron. Also, neurons transmit chemical signals known as neurotransmitters to other neurons, which have different functions depending on its physical composition.

Remember now the neural version of the replacement thought experiment. The replacement process is performed in such a way that all the members of the sequence <j1, … jn> are functionally equivalent to the original being (in this case, Paul´s brain). If it is shown that the last member of the sequence is effectively conscious, then what it is shown is that the properties that have been preserved by the replacement process are sufficient for generating conscious phenomenal states. However, it is not only the property of instantiating a certain functional organization that was preserved through the replacement process. Electrical and chemical signals are also transmitted by artificial neurons. Since this element is still present, we cannot conclude that consciousness is preserved in virtue of the

functional organization of the system: it might be still there because the signals transmitted by artificial neurons are the same ones that are transmitted by organic neurons.

At the end of the last chapter, we have seen that there is a way for describing the neural version of the replacement thought experiment without stipulating that the artificial neurons preserve the electrochemical signals generated by organic neurons. The problem is that, if there is no restriction concerning the materials on which the artificial neurons are made, and if the physiochemical constitution of the signals transmitted by organic neurons can be dispensed with, the functional organization that is implemented by the system may be shared by a very large number of entities. Note that there might be a huge number of physical systems in the universe whose number of parts is comparable to the number of neurons located inside a person's brain. These parts may also exhibit a number of physical states that can be put into correspondence to the physical states of a neuron, and they may show a certain level of causal interactions that can be described as the interchange of inputs and outputs, analogously as the interchange of electrochemical signals by neurons inside the brain. Among these systems there might be some ones whose functional organization at some level matches the functional organization that the brain of a person has at the neural level. If this is true, then it does not make sense to say that the human brain implements a certain functional organization, since we can take almost any physical system and interpret them as instantiating any functional organization, including, of course, the functional organization of the brain at a neural level.

In the first section of this chapter, I will present a similar problem mentioned by Block in (2007b). I also explain how this problem is a challenge for the replacement strategy. In the second section of this chapter, I will present an argument offered by Searle in (1980) in favour of what he calls the thesis of Universal Implementation. According to the reconstruction I will propose, the crucial premise in this argument is that computation is observer-relative: whether a system implements a computation or not depends on how an external observer interprets this system. In the third section of this chapter, I will discuss the conditions of implementation of a Combinatorial State Automaton (CSA) given by

Chalmers. We will see that, in spite of Searle´s objections, these conditions show that it is extremely unlikely that an arbitrary object implements a CSA.

Section 1
Chauvinism, liberalism and the replacement thought experiment

In (2007b), Block perceives a similar problem with respect to the specification of the inputs and outputs inside a functional theory of mental states. According to him, no matter how these inputs and outputs are physically described, the theory will be either chauvinistic – that is, it will deny mentality to other systems simply because they do not share the physicochemical composition of organic neurons – or too liberal – it would wrongly attribute mentality and the capacity of experiencing mental states to systems clearly unable to have it. The replacement thought experiment faces then the following dilemma. Either the input and output signals transmitted by artificial neurons are the same ones transmitted by organic ones, or they are simply described as inputs and outputs, without any specification concerning their physicochemical constitution. In the first case, the conclusion would be too chauvinistic, since "precludes organisms without neurons (e.g., machines) from having functional descriptions." In the second case, the risk is that, due to the liberal way in which the inputs and outputs are specified, a huge number of systems may implement the functional organization of the brain.

Block illustrates this risk by mentioning the possibility of manipulating the economic system of a country in such a way that it duplicates the functional organization of the brain at the neural level. To begin with, we may think of the inhabitants of this country as the constituent parts of the system, and the whole value of their assets as their relevant internal state. We may also take the money earned by these inhabitants as the input and their expenditures as the output.

> Economic systems have inputs and outputs, e.g., influx and outflux of credits and debits. And economic systems also have a rich variety of internal states, e.g., having a rate of increase of GNP equal to double the Prime Rate. It does not seem impossible that a wealthy sheik could gain control of

the economy of a small country, e.g., Bolivia, and manipulate its financial system to make it functionally equivalent to a person, e.g., himself.

Those who want to use the replacement thought experiment for arguing in favour of the thesis that implementing the right functional organization is sufficient for generating consciousness face a dilemma analogous to the one described by Block. Either the physicochemical composition of the input and output signals transmitted by the neural replacements is the same as the one of organic neurons, or the signals are described only as inputs and outputs, without any reference to their composition. If we adopt the first option, the replacement thought experiment and the arguments associated to it might succeed in showing that the ability to experience conscious phenomenal states is preserved, but it will be useless for arguing in favour of the thesis in question. The reason, of course, is that it is not only the functional organization of the brain what is preserved, but also the physiochemical character of the inputs and outputs interchanged by organic neurons. Thus, even if conscious experiences are preserved after the replacement process it is not possible to assure that this is in virtue of the system implementing the functional organization of the brain at a neural level. If we adopt the second option, the risk consists in that the number of systems that share the functional organization of the brain might be too large. It may be argued that there are millions of physical entities – rocks, heaps of sand, galaxies, bars of soap – that can be divided into the same number of parts as neurons inside the brain of a person. These parts maintain certain causal interactions that can be interpreted as signals. Some of these causal interactions might be isomorphic to the way in which neurons are causally affected among themselves. Stars inside galaxies, for instance, have certain causal effects on other celestial bodies. The same can also be said of molecules of a rock made of granite, or of grains of sand in a heap.

Section 2
The thesis of Universal Instantiation

Objections like the one presented by Block have been widely discussed in the philosophical literature during the last forty years. In particular, some philosophers have developed what are known as “triviality arguments” against computational approaches to mentality. Perhaps

one of the most well known arguments is the one presented by Searle in (1992), where this author offers an argument whose conclusion is the thesis of Universal Implementation, which can be initially understood as follows: there is always a way of interpreting an object in such a way that it implements any computer program. Remember that in the second chapter of the thesis we mentioned that any given functional organization could be abstracted into an abstract computational device: a Combinatorial State Automaton. But if the thesis of Universal Implementation is true, then there is always a way of interpreting an object as implementing a CSA, and thus, there is always a way of interpreting it as implementing any functional organization. In this section, I will present and evaluate a reconstruction of Searle´s argument, which includes a key premise: computation is observer relative. Searle argues that the question whether an object (which may be a digital computer, a rock, a wall or an organic brain) implements a certain program does not have an objective answer: this depends on the way an external observer interprets this object. The Thesis of Universal Implementation can be formulated as follows:

> (Universal Implementation) For any computer program C and any sufficiently complex physical object O, there is a description of O under which it is implementing program C.

Now, if the argument offered by Searle is successful, the consequence is that almost every physical object we choose (provided its structure is complex enough) can be interpreted as implementing any program. Remember the modified versions of the supervenience thesis (NSS_M) that was presented in page (47) of this thesis:

> (NSS_M) There is a set *C* of Combinatorial State Automata such that the property of a system *P* of generating conscious phenomenal experiences *naturally* supervenes on the property of *P* of instantiating a member of set *C*.

If thesis (NSS_M) is accepted and the thesis of Universal Instantiation is true, then almost any object can be described as implementing any CSA, and in particular, a CSA that corresponds to the functional organization of the brain at a neural level. Thus, the

unwelcomed consequence is that almost any object is capable of generating conscious phenomenal experiences. The claim that we can assign a computational interpretation to any object is understood by Searle as follows:

> There is no way you could discover that something is intrinsically a digital computer because the characterization of it as a digital computer is always relative to an observer who assigns a syntactical interpretation to the purely physical features of the system. […] As applied to the computational model generally, [this has the consequence that] the characterization of a process as computational is a characterization of a physical system from outside; and the identification of the process as computational does not identify an intrinsic feature of the physics; it is essentially an observer-relative characterization. [48]

From this, Searle obtains the following two theses: (1) for any object there is some description of that object such that under that description the object is a digital computer, and (2) For any program and any sufficiently complex object, there is some description of the object under which it is implementing the program.

Searle also mentions a famous example involving the molecular structure of a wall: according to him, there is a pattern of molecule movements inside this wall that is isomorphic with the formal structure of Wordstar (a word processor widely used in the 1980s). Now, provided that is a big enough wall, there will be a way of interpreting another pattern of molecule movements inside it as isomorphic to the formal structure of a different program, including, for instance, the formal pattern of neurons firing inside an organic brain.

In this section, I want to concentrate on the thesis of Universal Implementation. The reason is that, if true, this thesis is the one that presents a serious challenge to the reformulation of the supervenience thesis: there is a computer program C such that the generation of conscious phenomenal experiences by a system supervenes on the property of instantiating program C. If, on the one hand, we accept this thesis and on the other the thesis of Universal Implementation is true, then the obvious consequence is that any system

48" Searle (1992, p. 211)

implements program C and therefore, any system is capable of generating conscious phenomenal experiences. This, however, is against our basic intuitions: things like walls or rocks are not the sort of things that can generate these experiences.

The following is the reconstruction of the argument proposed by Searle in favour of the thesis of Universal Instantiation:

(Premise 1) Computer programs can be defined entirely as syntactic manipulations of symbols.

(Premise 2) Syntactical manipulations are not defined in physical terms.

(Conclusion 1) Computer programs are not defined in physical terms.

(Premise 3) Whether a physical system implements a computer program or not depends on an observer.

(Conclusion 2) For any computer program and any sufficiently complex object, there is some description under which such object is implementing the program.

In order to evaluate this argument, I want to start first with the truth of its premises. Consider first premise (1): a computer program is defined entirely as a syntactic manipulation of symbols. This can be understood with the help of the notion of a Turing Machine. A Turing Machine is an abstract device composed of a reading-writing head and an infinite tape divided into cells. The head is capable of scanning (one at a time) these cells and deleting and writing symbols on them. Also, the machine is, at any time, in one of a finite number of states. The way in which the machine operates is completely determined by (a) the current state of the machine, (b) the symbol currently scanned by the head, and (c) a set of transition rules, known as the program of the machine. The following is a program that defines a very simple Turing Machine, whose task is adding the symbol "1" to any chain of 1´s written on the tape:

1. [S_0, 1, S_0, right]
2. [S_0, 0, S_1, wrt1]

3. [S_1, 1, S_1, left]
4. [S_1, 0, S_2, right]

This Turing Machine can be understood as computing the function that assigns to any natural number its immediate successor (the successor function). This requires adopting a certain interpretation of the symbols that appear in the tape cells. In this case, we can represent any arbitrary natural number n as a series of n+1 instances of the symbol "1" written on the machine tape. The symbol "0" can be used for separating blocks composed by 1´s. For all blocks of length n, the machine will always end with a block of length n+1. For instance, if the symbols on the tape are seven 1´s (representing the number 6 in this particular interpretation) the tape of the machine will end with a block of eight 1´s (that represent the number 7). Under this interpretation, the machine can be understood as computing the successor function. Of course, it might have been possible to choose another way of representing natural numbers: in binary notation, for example, the number 2 can be represented with the block "10", instead of "11". In this case, we would need to build another program for calculating the same function. But although binary notation would provide some important practical advantages concerning computational advantages (computational speed and economy of space, to mention only a few) this would not imply a difference in the computational capability of the machine. With this in mind, it is easy to understand why computation is entirely defined as a syntactic manipulation of symbols. Any particular computation can be defined entirely by a machine table (like the Turing Machine that computes the successor function). The machine table determines how the symbols on the table are manipulated, and these rules do not consider any properties apart from the syntactic properties of these symbols.

According to the second premise of Searle´s argument, syntactical manipulations are not defined in physical terms. Consider again the successor function: we can construct a physical machine that computes this function for a finite number of arguments in the domain (since it is a physical machine, it cannot perform the computation for all possible elements of the domain). Notice that while a particular physical machine can perform calculations by writing symbols on a tape. An example could be a physical simulation of a

Turing Machine built for educational purposes. Other physical machines used for computing functions, however, do not need to perform the calculations in this same fashion. A physical simulation of a Turing Machine might include a real head that scans symbols on a small paper tape, and can use a little pencil for writing the necessary symbols. Abacuses and digital electronic calculators, however, will use beads sliding on wires or electronic signals. But in all these cases, the fact that these machines are computers is independent of the particular physical ways in which they perform the calculations. In this sense, hardware is irrelevant for defining how a computer works. This contrasts with the way in which we understood and define mechanical and organic devices: a carburettor, for example, is a device that mixes air and fuel. Hearts are muscles that pump blood through the blood vessels by repeated contractions. The definition of a computer, however, takes in consideration only the syntactical properties of symbols, which according to Searle, are not intrinsic to physics, insofar as they have no physical effects. This is the way in which Searle justifies the claim that syntactic manipulations are not defined in physical terms.

> The multiple realizability of computational equivalent processes in different physical media is not just a signt that the process are abstract, but that they are not intrinsic to the system at all. They depend on an interpretation from outside. (Searle, 1992, p. 209)

Consider again the reconstruction of Searle´s argument: the step from premises (1) and (2) to the first conclusion seems to be valid: since computer programs can be completely defined as syntactic manipulations of symbols and that syntactic manipulations of symbols are not defined in physical terms, we have to accept that computer programs are not defined in physical terms. The step to conclusion (2), however, is problematic: that the objects of a certain field or theory cannot be physically defined does not mean that only an external observer can define them. The solution might be to include an additional premise:

(P3) If something is not defined in terms of physical features, then it is observer-relative.

In the following paragraphs, I want to argue that the truth of this thesis is at least doubtful: that some features of the world could not be defined by appealing to physical properties does not entail that the usual definitions of such features depend on an observer.

Section 3

The conditions for implementing a computation

Notice, first, that there seems to be certain clear constraints concerning the way in which we can we interpret a physical device. Consider the following argument given by Block[49]: imagine a physical realization of a Turing Machine whose tape only includes symbols composed of 1´s and 0´s. When the symbols in the machine tape are (1,1), the machine will produce the symbol "1", and when the symbols in the tape are (1,0), (0,1) or (0,0), the machine will produce the symbol "0". The relation between inputs and outputs can be given by table 1:

Table 1

Input 1	Input 2	Output
1	1	1
1	0	0
0	1	0
0	0	0

The machine can be interpreted as instantiating the truth function "and" when the symbol "1" is interpreted as "T" or "True", and the symbol "0" is interpreted as "F" or "false", as it is shown in table 2.

Table 2

Input 1	Input 2	Output
T	T	T
T	F	F
F	T	F
F	F	F

49" Block (2002, pp. 77-78)

Note that it can also be interpreted as instantiating the truth-function "or" if we interpret the symbols "0" and "1" in an opposite way, as it is shown in table 3.

Table 3

Input 1	Input 2	Output
F	F	F
F	T	T
T	F	T
T	T	T

This particular Turing Machine can be interpreted in more than one way. However, consider now exclusive disjunction (table 4): no matter the meaning attributed to the input and output signals, the machine cannot be interpreted as instantiating this truth-function.

Table 4

Input 1	Input 2	Output
T	T	F
T	F	T
F	T	T
F	F	F

Also, it is evident that certain features of the world are observer-relative: that London springs are terrible for a picnic, that the taste of a Vesper cocktail is superior when you use Kina Lillet instead of Cocchi Americano, or that the sound of a Stradivarius violin is deeper than the sound of a Guarneri, depend on the point of view of an observer. But notice that although some features of the world cannot be definable in a physical theory does not mean that they are observer-relative. In (2002), Rey argues that some categories that belong to the empirical sciences are clearly objective, in spite of not being defined *exclusively* in terms to their physical features[50]. Consider, for instance, a basic category used in biological classification: the notion of *species*. An adequate definition of this notion is based on

50 Rey (2002, p. 215)

certain relational properties such interbreeding or the capability of produce new offspring, and not only on physical properties. This notion is not reducible to physics, but from this we cannot conclude that it is observer-relative.In some other cases, however, there is a debate concerning the nature of certain features. In (2008) Buechner mentions the case of numbers. He reminds us that their metaphysical nature is a debatable case, and it is not evident that they are not intrinsic to the world. Under a realist perspective, numbers are intrinsic features or some world. From a constructivist view, they can be seen as observer-relative.

> The point is that with respect to numbers, it is an open question whether numbers are or are not intrinsic features of the world or of some abstract world. If Searle thinks something is not an intrinsic feature of the physical world or an abstract world, it is a claim for which he must provide an argument. [51]

In spite of these initial considerations, I think that the most promising way for arguing against the thesis of Universal Instantiation is to show that the conditions under which a system implements a determinate CSA are sufficiently constrained for assuring that only a very limited number of systems implement it. The conditions of implementation are very important in fields like cognitive science, artificial intelligence and philosophy of mind, where we deal with the concrete implementation of abstract computations by concrete physical systems. Objects that implement a computation are concrete entities that follow physical laws and have certain particular constraints and limitations. For these reasons, it is necessary to explain how a determinate physical system implements or realizes a computation and to clarify the idea that a certain computation describes the operation of a particular physical system. Consider again the conditions proposed by Chalmers for implementing a CSA:

> A physical system implements a given CSA if there is a decomposition of its internal states into substates $[s_1, s_2, \ldots s_n]$and a mapping f from these substates onto corresponding formal states S^i of the CSA, along with similar mappings for inputs and outputs, such that: for every formal state transition $([I^j, \ldots I^k], [S^l, \ldots S^n]) \rightarrow ([S'^l, \ldots S'^k], [O^l, \ldots O^j])$ of the CSA, if the system is in internal state $[s^l, \ldots s^n$

51 " Buechnner, (2008, p. 160)

] and receiving input [i^1, ... i^n] such that the physical states and inputs map to the formal states and inputs, this causes it to enter an internal state and produce an output that map appropriately to the required formal state and oputut.[52]

The conditions established by Chalmers require that, for each formal state transition of a CSA, system P needs to satisfy a corresponding conditional that has the following form:

> If system P were to be in vector-state s and receives input i such that $f(s) = S$ and $f(i) = I$, then it would transit into a vector-state s' such that $f(s') = S'$ and would produce output o, such that $f(o) = O$.

Notice that conditionals that have this form possess a modal strength that is sufficient for supporting counterfactual conditionals: they establish what would be the case if their antecedent were true. This contrasts with simple material conditionals. The importance of these counterfactual conditionals is that their satisfaction by a system guarantees the reliability of its state transitions, and they assure that they are not produced by pure coincidence. Naturally, physical systems like brains or their artificial simulations might fail, but in order for implementing a CSA, it is not required that they transit adequately in all possible circumstances.

The strategy for showing that an arbitrary physical system, like Searle´s wall, cannot reliably implement any given CSA can be understood as follows: consider first a CSA_N whose vector-states have 10 elements that can be in 10 different states (for ease of exposition, assume that the sets Σ and Γ that correspond to the sets of inputs and outputs are empty). The number of possible state vectors of CSA_N will be 10^{10} (that is, 10,000,000,000 state-vectors). Now, if it were true that Searle´s wall can implement CSA_N, it will be necessary to find first an adequate mapping from the components of this wall to the corresponding vector-states of CSA_N. To each of these vector-states can be assigned, for instance, a small region of the wall. We can also identify a certain physical state-transition

52 " Chalmers, (1996b, p. 325)

in these small regions that correspond to the state-transitions of the CSA: they can change, for example, their colour or temperature.

However, this procedure will radically reduce the number of systems that mirror the vector-states of CSA_N. The reason is that these small regions of the wall need to satisfy the strong conditionals that describe the transition rules of CSA_N (in this case, remember that the sets of inputs and outputs are empty): if a physical system B is in vector-state $[s^i_1, \ldots s^i_{10}]$, it will transit into vector-state $[s^{i'}_1, \ldots s^{i'}_{10}]$. Notice that for each vector state we can have 10^{10} different consequents, and thus, the number of the transition rules of CSA_N can be as high as $10^{10} \times 10^{10}$. Now, if the transition between the states of the regions of Searle´s wall mirrored the corresponding vector-states of CSA_N, it is necessary for it to satisfy the corresponding transition rules. However, it is extremely unlikely that the regions of the wall transit in a way that mirrors these vector states. The wall would need to satisfy $10^{10} \times 10^{10}$ counterfactual conditionals, one for each transition rule. It is for these reasons that the possibilities that an arbitrary physical system implements CSA_N are minimal.[53]

Conclusions

According to the thesis of Universal Implementation, for any computer program *C* and any sufficiently complex physical object *O*, there is a description of *O* under which it is implementing program *C*. The conclusion of the replacement argument is that implementing the right functional organization is a sufficient condition for consciousness. But if any sufficiently complex object (say, an object with an adequate number of identifiable physical states, like a wall) implements any program, it would also implement the right program, and therefore any sufficient complex object would be conscious. We have seen in this chapter that, to be valid, Searle's argument needs to adopt an additional premise: if something is not defined in terms of physical features, then it is observer-relative. However, there are scientific notions that do not depend on an observer but are not reducible to physical features, like the notion of species. Moreover, there are clear

53 The situation may be even worse: remember that the sets Σ and Γ that correspond to the sets of inputs and outputs of CSA_N are empty. The number of the transition rules for implementing a CSA with inputs and outputs will add more constraints to the conditions of implementation.

constraints on the way that we can interpret a physical system: Block's argument shows that a system that instantiates conjunction or disjunction cannot be interpreted as instantiating material equivalence or exclusive disjunction.

But the main problem with the thesis of Universal Instantiation is that the implementation conditions of an abstract computational device, like a Combinatorial State Automaton, do not allow the trivial implementations suggested by Searle. Remember that the vector-states of a Combinatorial State Automaton are complex structures. In order to implement a CSA, it would be necessary for a physical system, like Searle´s wall, to satisfy the transition rules of the CSA that are expressed as counterfactual conditionals. But as we have seen, the possibilities that an arbitrary physical system satisfied these transition rules are minimal in the case of an inputless CSA with 10 elements and 10 states for each element. A simple mapping from the components of this wall to the vector-states of the CSA would not be sufficient. Moreover, there will be more constraints if we consider a CSA capable of dealing with inputs and outputs.

Chapter 4
The replacement thought experiment and the thesis of inverted qualia

Introduction

Although the version of the replacement thought experiment presented in the second chapter of this thesis could be adequate for showing that a replacement process that preserves the functional organization of the brain also preserves its capacity of generating conscious phenomenal experiences, there is still a further problem that needs to be faced. In the case examined in the last section, we have seen that there are very good reasons to think that a being that possesses an artificial brain that shares the functional organization of the brain of a conscious person will also be able to have conscious phenomenal experiences, no matter the particular physiochemical makeup of this artificial brain (leaving aside, of course, chemical substances that may be harmful to the rest of the brain or to the body). More precisely, the arguments presented in the last section of this chapter support thesis (NSS):

> (NSS) There is a set of functional organizations F such that the capacity of a physical system of generating conscious phenomenal experiences naturally strongly supervenes on the property of P of instantiating a member of set F.

However, it still can be argued that the experiences that this physical system generates have a different character than the experiences generated by a biological brain. According to this, it is possible that a being that possess an artificial brain made from non-biological basic elements but that instantiates the same functional organization of the brain of a conscious person might experience pain, smells or sounds in a different way in which a being possessing an organic brain experiences pain, smells or sounds. The artificial brain of Paul's functional twin, F-Paul, may have the capacity of generating conscious experiences, but they still may be different if they are compared with the experiences generated by Paul's organic brain. Thus, it is still necessary to determine whether the replacement thought experiment presented in the preceding section of this chapter supports thesis (NSS_2):

(NSS_2) There is a set F of functional organizations such that the qualitative character of the conscious phenomenal experiences of a physical system supervenes naturally on the property of P of instantiating a member of set F.

A widely suggested way of explaining this possibility is to accept that the mere presence of conscious phenomenal experiences is generated by the functional organization of the brain, while the particular character of these experiences depends not on this functional organization, but on the particular physicochemical makeup of the system. Systems that share the same functional organization but whose physicochemical composition is different will generate conscious phenomenal experiences with different qualitative character. Assuming this claim, it is not difficult to see why the conscious phenomenal experiences of Paul and F-Paul might be different. Their respective brains (a biological brain and an artificial brain) share the same functional organization at a very fine-grained level, and it is for this reason that these brains generate conscious states. But Paul´s brain is composed of organic matter, and this particular composition determines the qualitative character of his conscious experiences, a character that is not shared by the conscious experiences generated by F-Paul´s artificial brain, which is not composed not of organic matter. In a hypothetic case in which a being could have the conscious experiences generated both by Paul´s brain and by F-Paul´s brain, the qualitative differences between them could be noticed.

In the first section of this chapter, I will present the hypothesis of inverted qualia and the challenges it presents to a functionalist account of phenomenal consciousness. In the second section, I will present a further variation of the replacement scenario that supports thesis (NSS_2). According to this thesis, the phenomenal character of the experiences generated by a system naturally supervenes on the property of instantiating a certain functional organization. In this case, the functional organization at play is the functional organization of the brain at a neural level. If thesis (NSS_2) is true, the inverted qualia hypothesis is naturally false: it is not the case that a system that is a functional duplicate of the brain of a conscious person at a neural level generates conscious phenomenal

experiences with a different qualitative character. In the third section of this chapter, I will present an objection against this new version of the replacement thought experiment originally developed by Van Heuveln *et al* (1998) and Greenberg (1998). Finally, in the fourth section of this chapter, I will present a response against the objections raised by Van Heuveln *et al*.

Section 1

The hypothesis of inverted qualia

How can be understood the claim that functionally identical states may have different qualitative characters? There are surely several ways of imagining the nature of this difference. Consider the case of pain. From a functionalist point of view, pain is an internal state that can be characterized by its relations to other internal states, perceptual inputs and behavioural outputs. A subject is in pain when he is experiencing a state that is normally caused by an injury to the body, that produced the belief that the body has been damaged and the desire to be out that state, and that in general produces screams, moaning and other expressions of discomfort. Imagine now that Paul and F-Paul find themselves in an identical situation: they are in a scenario rehearsing a song for a concert. While they are busy preparing the audio equipment and tuning the instruments, they grab a wire lying in the floor and try to connect a guitar to one of the amplifiers. At this precise moment, they receive a painful (but harmless) electrical discharge. Both of them experience an uncomfortable sensation. They close their eyes, scream, and let the wire fall on the floor. This experience generates in both of them the distant memory of having received a similar electrical discharge long time ago (perhaps during the initial rehearsals of the Beatles in Liverpool), and the belief that there must be something wrong with the wire. The incident also causes them to angrily order the staff to replace it as soon as possible in order to avoid a more severe accident. Now, in spite of the fact that these experiences are caused by the same electrical discharge, that they cause the same beliefs and memories, and produce the same behavioural effects, the possibility just discussed implies that Paul and F-Paul painful experiences might have had a very different qualitative character, a difference that could be perceived by a single individual in case that he had these experiences.

However, I think it is difficult to coherently imagine a case in which the conscious phenomenal experience of pain can be different between two isomorphic systems. Assuming that both Paul and F-Paul experience something uncomfortable when they receive an electrical discharge, and that their cognitive and behavioural reactions are identical, it is not clear how these experiences could be dissimilar. Perhaps the painful experience of F-Paul is comparatively weaker than the experience of Paul, but their reactions to the electrical discharge are the same because, after all, their motor neurons send the same signals to the muscles of their bodies. However, it would not be clear then why their cognitive reactions to these experiences are identical. Perhaps F-Paul experiences a radically different sensation, and that any being possessing an organic brain would be amazed if he could feel the painful experiences of F-Paul.

Leaving aside the case of pain, there is a way of conceiving a difference in the qualitative character of the experiences generated by isomorphic systems: the hypothesis of the inverted spectrum. According to this hypothesis, the visual experiences generated by two isomorphic systems, P and Q, may differ in that the qualitative properties of the visual experiences generated by Q are phenomenally inverted with respect to the visual experiences generated by P. For example, Paul might be driving a car along Oxford Street, and noticing that the traffic signal turns to red at the intersection with Regent´s Street, he stops his car. The experience of seeing the change in the traffic light might also cause in Paul the belief that he must stop his car in case he does not want to be involved in an accident. Also, this experience might produce an uncomfortable sensation of anxiety, since he is late for a meeting and knows that he must wait at least four minutes in that busy intersection. When the four minutes have passed, Paul sees that the traffic light is now green and he continues driving. In an identical situation, Paul´s functional isomorph, F-Paul, would react in exactly the same way. He would stop his car when the traffic light changes its colour. He would also form the belief that, in order to avoid an accident, he must stop his car. He would also experience the same feeling of anxiety. But if the visual experiences of F-Paul were inverted with respect to the visual experiences of Paul, he would not notice that the traffic signal changes from green to red when he is driving along

Oxford Street. For him, the signal transits from red to green, and it is precisely when he sees the colour green that he stops his car. Of course, he does not call that colour "green". He is functionally identical to Paul, so his behaviour – including his verbal behaviour – is identical to the one exhibited by Paul.

In the rest of this chapter, I will present a version of the replacement thought experiment whose objective is to show that the inverted qualia hypothesis is false: the conscious phenomenal experiences generated by two isomorphic systems – like the organic brain of Paul and the artificial brain of F-Paul – cannot differ in their qualitative character. If the thought experiment is sound, then we have very good reasons to accept thesis (NSS_2): there is a set F of functional organizations such that the qualitative character of the conscious phenomenal experiences of a physical system supervenes naturally on the property of P of instantiating a member of set F.

Section 2

Against the thesis of inverted qualia

If the inverted qualia hypothesis is true, then there can be two isomorphic physical systems – like the organic brain of Paul and the artificial brain of F-Paul – that generate visual experiences whose qualitative character is mutually inverted. Paul sees the traffic signal red and a blue sky, while F-Paul sees the signal green and a yellow sky. Assuming that Paul and F-Paul are the extreme members of a sequence $< J_0, J_1, J_2, , \ldots , J_n >$ of replacement cases, the qualitative character of the visual experiences of these isomorphic systems will change gradually between the red experiences of Paul and the green experiences of F-Paul. Notice that the hypothesis of a sudden change in the qualitative character of these experiences (a change caused by the replacement of a single organic neuron) can be ruled out by appealing to the same reasons presented in the previous chapter. It is extremely unlikely that the replacement of only one neuron can have this effect in the experiences of the subject.

In this new version of the thought experiment, the hypothesis is that the qualitative character of these experiences gradually changes as the organic neurons of the brain are

being replaced by artificial neurons that preserve the same input-output relations with the rest of the brain, while the functional organization of the whole system remains constant. If the inverted qualia hypothesis is accepted, perhaps these experiences transit from red to green and from blue to yellow, or maybe the way in which these experiences change is so radically different that they would be amazed if they could interchange these experiences (maybe expressing this amazement by saying things like "It´s surprising! I´ve never seen this colour before!"). This qualitative difference would not be manifested in the behaviour of the subject because the envisaged scenario preserves all the relations existing among colours.

Now, assuming the truth of the inverted qualia hypothesis, there will be two points J_i and J_n in the sequence such that they are not the extreme points and that the qualitative character of their visual experiences is so different that an independent observer would notice it. It is evident that, as the extremes of the replacement sequence, there would be a noticeable difference between the red experiences of Paul and the green experiences of F-Paul. Nonetheless, there will surely be another pair of elements of the sequence that are closer between them, and whose visual experiences are also noticeably distinct. For ease of exposition, we will call this elements Zero-Paul and Ten-Paul. The difference between them consists in that Ten-Paul has ten per cent more artificial neurons in his brain. But how can we be sure that the visual experiences between Zero-Paul and Ten-Paul are noticeably different? Certainly, if there is a difference, it would not be as strong as the difference between the visual experiences of Paul and F-Paul, since the places in the sequence occupied by Zero-Paul and Ten-Paul are relatively closer. It even might be suggested that the differences in his visual experiences are so faint that an external observer would not notice any difference between them. Notice, however, that it is not possible to transit from red to blue with ten imperceptible changes, so we can expect that the pairs whose difference is ten per cent of artificial neurons have noticeable different visual experiences. If this is so, the difference concerning the qualitative character of the experiences of Zero-Paul and Ten-Paul will be noticeable, although this difference would not be as radical as the difference between the experiences of Paul and F-Paul. The importance of this point will be evident in the last section of this chapter, when we consider a response to an objection (presented by

Greenberg (1998) and Van Heuveln *et al* (1998) against this new version of the replacement thought experiment.

Now, Zero-Paul and Ten-Paul are members of the sequence $< J_0, J_1, J_2, , \ldots , J_n >$ of replacement cases, and thus, they are functionally identical (and almost physically identical). The only difference between them is that there is a section of the brain of Ten-Paul that has been replaced by artificial neurons, while that section remains organic in the case of Zero-Paul. Call the section inside the brain of Ten-Paul that is composed by artificial neurons the section *A*, and the corresponding organic section inside Zero-Paul the section *O*. Assume now that section *A* is removed from the brain of Ten-Paul and implanted inside the brain of Zero-Paul in such a way that, when a switch is in position *a*, section *A* is connected to the brain of Zero-Paul, and when the switch is in position *o*, section *O* is. Now, if the Inverted Qualia hypothesis were right, then there would be a noticeable change in the qualitative character of the visual experiences of Zero-Paul when the switch is moved from position *o* to position *a*. The visual experiences of Zero-Paul would change, in front of his eyes, from a reddish experience to a greenish one. At this point, we can form the following two hypotheses: either Zero-Paul notices that his visual experiences are changing in front of him, or he does not.

Does Zero-Paul notice that his visual experiences are changing in front of his eyes? Note first that, like in the case presented in the preceding chapter of the thesis, the behavioural dispositions of Zero-Paul will be the same before and after the switch is moved from position *a* to position *o*. The reason is that the causal effects of section *A* – composed entirely by artificial neurons – are the same causal effects of region *O*. Both regions will send the same signals to the rest of the brain, and in particular, they will affect the motor neurons in exactly the same way. Thus, the behaviour of Zero-Paul will be consistent with the preservation of the qualitative character of his visual experiences. If somebody asked him "Did you noticed any difference in your visual experiences after the switch was moved?" he would answer negatively.

> ... there is no way for the system to *notice* the changes. Its causal organization stays constant, so that all of its functional states and behavioral dispositions stay fixed. As far as the system is concerned,

> nothing unusual has happened. There is no room for the thought "Hmm! Something strange just happened!" In general, the structure of any such thought must be reflected in processing, but the structure of processing remains constant here. If there were to be such a thought, it must float entirely free of the system and would be utterly impotent to affect later processing. (If it affected later processing, the systems would be functionally distinct, contrary to the hypothesis). [54]

Notice also that the functional organization of Zero-Paul is exactly the same before and after the switch is moved, and thus, the way in which his brain processes information will not be modified. This makes extremely unlikely that he can form the belief that there is something different about his visual experiences when the switch is moved from position *o* to position *a*.

> It seems entirely implausible to suppose that my experiences could change in such a significant way, with my paying full attention to them, without my being able to notice the change. It would suggest once again a radical dissociation between consciousness and cognition. If this kind of thing could happen, then psychology and phenomenology would be radically out of step; much further out of step than even the fading qualia scenario would imply.[55]

Thus, Zero-Paul cannot form the belief that there is something different or strange about these experiences. But to assume this implies that there is a serious incompatibility between his conscious experiences and the beliefs he forms about them. Zero-Paul is a rational being that pays attention to his experiences, but he is simply not able to notice the change produced by the movement of the switch. To assume that he judges that nothing strange is happening to his visual experiences while they are changing in front of his eyes clashes with the Coherence Principles proposed by Chalmers. In particular, there is a conflict with this assumption and the Reliability Principle. According to this assumption, the second-order judgements made by the subject are normally correct. So, if Zero-Paul judges that he experiences pain in his right toe, or that he hears the sound of a bass guitar, he usually has the experience of pain in his right toe and the experience of hearing a bass guitar. Similarly, if he forms the judgement that his visual experiences are not changing in

54 " Chalmers (2010, p. 24)

55 " Chalmers (1996, p. 269)

front of his eyes, we would expect that this judgement is accurate. But as we have seen, this is not the case. His experiences are changing in front of his eyes, and he is simply unable to form the judgement that something strange is happening. Again, the clash between the Coherence Principles and the fact that Paul cannot form the belief that something wrong is happening with his conscious experiences cannot be interpreted as a genuine *reductio*. The Coherence Principles are not logically necessary, and there is no contradiction in assuming that they can be false in some circumstances. This version of the replacement argument, thus, does not show that the qualitative character of the conscious phenomenal experiences logically supervenes on the property of instantiating a certain functional organization. However, it supports thesis (SN_2): There is a set F of functional organizations such that the qualitative character of the conscious phenomenal experiences of a physical system supervenes naturally on the property of P of instantiating a member of set F. In this case, the relevant functional organization will be, again, the one instantiated by the brain at a neural level.

Section 3

Change of experience and the replacement thought experiment

Van Heuveln *et al* (1998) and Greenberg (1998) raise similar objections against this new version of the replacement scenario. The problem detected by them is that there seems to be an unjustified assumption concerning the identity of the subject whose brain suffers the replacement. According to their criticism, it is not possible to guarantee that there is one and the same individual before and after such replacement. It is perfectly possible to argue that the replacement process generates two different beings, each of them having its own phenomenal world. In such a case, it does not make sense to say that a single individual could have noticed a change in the qualitative character of the visual experiences, since the experiences before and after the replacement are owned by two different beings.

The criticism made by Van Heuveln *et al* has two different parts: the first consists in showing that the dancing qualia argument is invalid: there is a gap in the argument that consists in that to say that a single physical system is able to experience and to report a

change in the qualitative properties of his visual experiences is not justified. As far as the argument is presented, it is perfectly possible to say that there are two different physical systems and that, even if they have different qualitative experiences, it does not mean that there is a single system able to experience this change. This gap makes the dancing qualia argument invalid. The second part consists in showing that this gap cannot be repaired. According to Van Heuveln *et al*, the change in the qualitative properties of the visual experiences of the physical system cannot itself be experienced. Van Heuveln *et al* present two further arguments whose objective is to show that the argument cannot be fixed. As Chalmers presents the dancing qualia argument, there is an incompatibility between the claim that the subject – in our case, Zero-Paul– is able to notice the change in the qualitative character of his visual experiences, and the fact that his behaviour and all his cognitive states are preserved – by hypothesis – after the replacement. What Van Heuveln *et al* try to show is that it is perfectly possible to imagine a situation like the one envisaged by the replacement scenario of the dancing qualia argument in which there is a qualitative change but there is no clash or incompatibility with the behaviour and the cognitive states of the subject.

The claim made by Van Heuveln *et al* is that the replacement scenario on which this version of the thought experiment is based leaves an open gap that makes the argumentative strategy invalid. Remember that part of the strategy of the argument is to show that there is a clash between the Coherence Principle and the assumption that the qualitative character of the visual experiences of the subject change when the switch is moved from position *o* to position *a*, and as a result, there is a change in the visual experiences of the subject. However, Van Heuveln *et al* claim there is no justified reason to think that the movement of the switch preserves a single individual able to perceive this change. When the switch is moved, there are two possible outcomes. The first consists is that when the switch is moved from position *o* to position *a*, one and only one individual experiences a change in the qualitative character of his visual perceptions. The second is that flipping the switch marks a boundary between two different individuals whose conscious phenomenal experiences are distinct. Thus, it is perfectly possible to imagine that changing the position of the switch generates different, independent individuals having their own conscious phenomenal

experiences. Furthermore, the fact that there are in fact two different physical systems involved (they are not identical: one of them has ten per cent more artificial neurons than the other) seems to justify the second outcome. But if this is true, there would be no problem in thinking that the qualitative character of the visual experiences is modified by moving the switch: since there is not a single individual, no one would notice a change in the qualitative character of the visual experiences, and thus, there will not be any incompatibility between the Inverted Qualia hypothesis and the Coherence Principle.

This same problem is also mentioned in Greenberg (1998). According to him, the assumption consists in that the system can meaningfully remember and compare qualitative conscious experiences, like the state of perceiving a red traffic signal when the switch is in position *o* and the subsequent state of perceiving a green traffic signal when the switch is in position *a*. The problem is the assumption that there is a unique physical system involved: when the position of the switch changes, we have a different physical brain. Greenberg admits that the sense of personal identity may not change when the switch is moved. By the assumptions of the thought experiment, we know that the subject would not say that he perceived something that makes us think that his personal identity was modified. He would still report that he has a continuous temporal perception of a green field. Nonetheless, this does not assure us that these two brains – the one that works when the switch is in position *o*, and the one that works when the switch is in position *a* – could meaningfully compare their qualitative visual experiences.

> ... no matter what the size of the switched module, there are still two different instantiations of brain/minds in this thought experiment. It is certainly possible that the experienced *sense* of personal identity may not change: there might be a report of a continuous sense of self experienced between the two alternating instantiations, an interesting state of affairs in its own right (it should be noted that a continuous experienced sense of self associated with even one physically instantiated brain is an elaborated construct). However, it does not follow that the two brains would experience colour in the same way, or could meaningfully compare the qualitative experiences of one with the other (unless one already assumed the principle of organizational invariance).[56]

56 " Greenberg, 1998, p. 55

What Greenberg says is that that there is an unjustified assumption behind this version of the replacement scenario. The idea is that there is something like an "inner eye", which binds visual experiences together and can make sense to the claim that these visual experiences are perceived by a single self.

The strategy adopted by Van Heuveln *et al* consists in arguing in favour of the idea that the replacement scenario can be viewed differently. Consider now how they reconstruct this scenario: if we remember the dancing qualia thought experiment that was examined in the second chapter of this thesis, there are two elements J_i and J_n that belong to the sequence of cases $< J_0, J_1, J_2, , \ldots , J_n >$ such that the qualitative character of their visual experiences is different. These two elements are Zero-Paul and Ten-Paul. We also identified the region of the brain that makes Zero-Paul and Ten-Paul different physical systems, which is precisely the region that is composed by ten per cent more artificial neurons: the *A* region. The next step was to imagine that the *A* region was installed in Zero-Paul´s brain alongside the equivalent biological region, and that there was a switch that connects, alternatively, these two different regions to the rest of Zero-Paul's brain.

The point of Van Heuveln *et al* is that there are two clearly different physical systems when the switch is moved: again, they cannot be identical since one of them has ten per cent more artificial neurons than the other. Lets understand the expression "connected system" as denoting the physical system that is connected to the rest of the brain. In this way, we will say that when the switch is in position *O*, the connected system is Zero-Paul, and when the switch is in position *A* the connected system is Ten-Paul. When Zero-Paul is the connected system, the reports concerning conscious phenomenal experiences are sent by Zero-Paul, while the reports are being sent by Ten-Paul when it is the connected system. Once the scenario is described in this way, we might ask whether the new connected system experiences a change of experience or not when the switch is flipped.

Imagine now that Zero-Paul is looking to the blue eyes of Linda McCartney in an old photograph. According to the reconstruction of the scenario proposed by Van Heuveln

et al, all the reports of conscious experiences are produced, at this moment, by Zero-Paul. Suppose now that the switch is moved: section *O* is disconnected and its place is taken by section *A*. From this moment, the reports of the experiences are sent by the new connected system, that is, by Ten-Paul. Assume now that the visual experiences of the new system, Ten-Paul, which is situated in the same perceptual circumstances as Zero-Paul, are different. Thus, the new connected system is experiencing the colour of Linda´s eyes as yellowish.

At this point, Van Heuveln *et al* formulate what they call the Crucial Question: when the switch is flipped, does the newly connected system experience any change? Remember that we have two different physical systems with the same functional organization: Zero-Paul and Ten-Paul. The question is whether one of these systems will experience a change when the position of the switch is modified. Since Van Heuveln *et al* understand Chalmer's argument as a *reductio*, they claim that, in order to formulate the *reductio* hypothesis, the answer to the Crucial Question should be positive.

> The Crucial Question provides us with an objective, safe, non-suggestive, and neutral question, because it is phrased in a way we can all agree on: It asks whether any of the two involved systems, having an identical functional organization but a different physical realization, will experience any change when the switch is flipped. […] In order for the dancing qualia argument to go through. Chalmers must answer the Crucial Question 'Yes'. [57]

What Van Heuveln e*t al* are claiming is that there is no guarantee that this experience of change happened. When the switch is moved and Ten-Paul becomes the connected system, we can perfectly imagine that this new connected system has visual experiences completely different from the ones had by Zero-Paul without assuming that this change of visual experiences was itself experienced as a change.

> The step from the change of experience to the experience of change is a non-trivial one and needs to be argued for. The gap in Chalmers' argument is that no such argument is being made. Instead, Chalmers do not notice the step, and simply equates the two. He thus makes an *equivocation fallacy* between *a*

57 " Van Heuveln *et al*, (1998, p. 243)

> *change in experience between two systems* and *a system's experience of change*, or between *some experiencing changing between two systems* and *some system experiencing change*.[58]

That there is a change in experience is, for Van Heuvelnt *et al*, uncontroversial. When the switch is flipped, there is a new connected system, Ten-Paul, whose visual experiences are inverted with respect to those of the former connected system, Zero-Paul. But a change in experience does not imply that any of these two systems had an experience of change.

> … in the thought experiment, the two systems alternate interacting with the same envoronment and share a lot of physical material. Thus one is being lured in to thinking that the change of experience equals an experience of change. However, just because some system's experiences differs from that of some other system does not mean that that system (or any system) experiences this difference. [59]

These are, in general terms, the reasons that Van Heuveln *et al* offer in favour of the claim that the dancing qualia argument is invalid.

Of course, that there are two physical systems does not need to be in conflict with the claim that there is only one individual. Paul is the same individual he was when he was five years old. Nonetheless, it is clear that Paul at five and Paul at sixty-four are different physical systems: the matter that constitutes his body is continuously renovated. In the same way, Zero-Paul is physically different from the being that results from flipping the switch to the E position.

A first problem with Van Heuveln *et al* objections consists in that it is doubtful that the replacement of a single module in the brain suffices for destroying the personal identity of the subject. It is true that Van Heuveln *et al* refuse to board the problem in terms of the notion of an individual. Nonetheless, we may perfectly ask whether such replacement has such catastrophic effect in the personal identity of the subject. After all, we can perfectly imagine the existence of an artificial device performing the function of some brain regions

58 " Van Heuveln *et al*, (1998, p. 244)

59 " Van Heuveln *et al*, (1998, p. 244)

without altering the personal identity of the subject. Greenberg also mentions that there are two different "brain preparations" involved in the replacement scenario. When the switch is moved, the perceived sense of psychological continuity may not be disturbed: Zero-Paul did not notice anything wrong with respect to his memories. Van Heuveln *et al* warn against evaluating their objections to the dancing qualia argument in terms of the criteria that make an individual an individual, and suggest that this might not be the wisest strategy. Although they do not make this point explicit, the argument of Van Heuveln *et al* against the dancing qualia argument can be perfectly understood without the claim that there are two different individuals involved.

Before presenting the first argument offered by Van Heuveln *et al*, it is important to mention that they adopt the following assumption: according to them, "It is plausible to assume that a system can only experience a change in his visual experiences through some recollection of previous visual experiences"[60]. As we will see, at least concerning visual experiences, this assumption is false. In the next section, I will present a version of the replacement thought experiment (presented originally by Zuboff in (1994)) that clearly illustrates a situation in which a change of the visual experiences of the subject does not involve any previous memories of such experiences. But leaving aside this assumption for a moment, Van Heuvlen *et al* suggest that there are at least two ways in which the recollection of visual experiences can be conceived. The first way is to imagine that the memories of past visual experiences had by Zero-Paul and Ten-Paul, together with their qualitative character, are stored in some part of the brain that functions like a databank. Recollection can be understood as a process in which memories of these visual experiences are picked up from this databank and then used for comparing the qualitative character of current visual experiences. In this way, it may be possible to notice the difference required by the dancing qualia argument and to answer "Yes" to the crucial question.

The second option proposed by Van Heuveln *et al* consists in that the process of recollecting past experiences uses the same psychological mechanisms required for generating the original visual experience. Furthermore, these mechanisms are part of the

60 " Van Heuveln *et al* (1998, p. 245)

region of the brain that is replaced in the dancing qualia thought experiment (that is, the neurons located in the visual cortex), since it is precisely because this part is replaced that the qualitative experiences of Zero-Paul and Ten-Paul are supposed to be different. Note that, if this were true, not only the qualitative character of the current visual experiences would be dependent on the physical nature of this region: this dependence would also affect the qualitative character of the past visual experiences recollected by the subject.

> On our view, recollecting visual experiences involves the use of some of the same mechanisms that are also involved when having current visual experiences. Moreover, these mechanisms with which the visual experiences are associated are, by the thought experiment, assumed to be part of the visual cortex. Therefore, the qualitative nature of these recollected visual experiences are just as dependent on the physical nature of the visual cortex in use as the qualitative nature of current visual experiences.[61]

Thus, it is perfectly possible to imagine that the qualitative character of Zero-Paul and Ten-Paul visual experiences is different: Zero-Paul may perfectly have green visual experiences when he is looking Granny Smith apples, or blue experiences of a deep summer sky, while Ten-Paul, who shares the same functional organization, has red visual experiences under the same circumstances. Nonetheless, neither of them would be aware of any change concerning the qualitative character of these experiences. The reason is not because there is a clash with the Coherence Principle, but simply because their recollections are generated by the same mechanisms that generate their current visual experiences. Zero-Paul would not only experience green Granny Smith apples and red strawberries: his recollections of past visual experiences would also be of green apples and red strawberries. In contrast, Ten-Paul recollections of past visual experiences would be of red Granny smith apples and green strawberries, which have the same qualitative character as his current visual experiences. Consequently, Zero-Paul will not notice any change when the switch is flipped. The argument can be reconstructed as follows:

> (Premise 1) Recollecting visual experiences involves the use of some of the same mechanisms that are also involved when having current visual experiences.

61 " Vah Heuveln *et al*, (1998, p. 246)

(Premise 2) These mechanisms are assumed to be part of the visual cortex.

(Conclusion 1) The qualitative nature of these recollected visual experiences is just as dependent on the physical nature of the visual cortex in use as the qualitative nature of current visual experiences.

(Premise 3) Although the qualitative nature of the experiences may be different between the two systems, the qualitative nature of the experiences that Zero-Paul remembers is similar to the qualitative nature of the experiences that Zero-Paul is having now.

(Conclusion 2) Zero-Paul will not experience any change in his visual experiences, and the fact that he reports no change in his visual experiences will not be surprising. [62]

The second argument presented by Van Heuveln *et al* is based on a further modification of the replacement scenario of the original thought experiment presented by Chalmers. Van Heuveln *et al* make the two following claims about this scenario: the first is that it is similar enough to the one presented in the dancing qualia argument for validating the same responses to all the relevant questions. The second is that it is different enough for allowing the reader to avoid the mistaken conclusions motivated by the way that the replacement scenario of the original argument was described. The modification of the replacement scenario proposed by Van Heuveln *et al* has two features:

(1) The first is that it is not only a small percentage of the brain that is replaced (for instance, the replacement part is made not only in the part of the brain responsible for the processing of visual experiences and the memories we have of them). Instead, the whole brain is substituted by a physical system with the same functional organization but made entirely of artificial neurons.

(2) The second feature is that the original biological brain is not preserved after the replacement. In contrast to the original replacement on which the dancing qualia argument is based, when a button is pressed the original brain gets destroyed and the artificial brain now occupies its place.

62 " Cf. Van Heuveln *et al*, (1998, p. 246)

Suppose now – as in the original replacement scenario – that the biological brain and the artificial brain differ in that they give rise to different qualitative visual experiences: a being with the biological brain would have a green experience when he looks the field in the park, while a being with the artificial brain would have a red experience in the same circumstances. When the switch is flipped, we will have a being that is almost identical to Paul. It has his same functional organization and behaves like him. The question is, does he notices a change in experience? According to Van Heuven *et al*, the answer is no. We can even imagine that each time the button is pressed a new brain functionally equivalent to the former one is created. This makes the existence of the new artificial brain completely independent of the previous existence of the biological brain. Thus, just because there was a biological brain before the button was pressed does not mean that the individual that has the new artificial brain notices a change in the qualitative character of his visual experiences.

The replacement scenario envisaged by Van Heuveln *et al* can have a further modification: we may assume that it is not only the biological brain that gets destroyed when the button is pressed, but Paul´s entire body. After the button is pressed, a being almost identical to Paul – with the only difference that it possesses an artificial brain – is created in the same place formerly occupied by him. As before, the functional organization of this new being is identical to Paul´s. They would also behave identically under the same circumstances.

Section 4

Visual memories and the replacement of the visual cortex

It is evident that the replacement of the whole brain of the subject by a functionally isomorphic artificial one does not warrant the preservation of the same individual. This is precisely what the modification of the thought experiment proposed by Van Heuveln *et al* shows. But the thought experiment offered by Chalmers clearly involves the replacement of a relatively smaller area of the brain. Remember that the justification for considering a replacement involving only ten per cent of the organic brain was that a change between the

visual experiences of Paul and the qualitatively inverted experiences of F-Paul could not be reached by ten unnoticeable changes. Thus, if the inverted qualia hypothesis were true, two members of the replacement sequence like Zero-Paul and Ten-Paul (whose respective brains differ in that one of them is composed by ten per cent more artificial neurons) would exhibit a noticeable difference in the qualitative character of their visual experiences. Since the replaced part of the brain is relatively small, it seems very dubious that a replacement like the one described marks the boundary between two different individuals. Of course, Van Heuveln *et al* insist that their argument does not rely on the fact that the replacement brings with it the generation of a new individual. Instead, they argue that recollecting visual experiences involves the use of some of the same mechanisms that are also involved in the production of these visual experiences. Thus, if these mechanisms were replaced by a system with different physical properties (like an artificial visual cortex) there is no guarantee that this system preserves these memories. So, that the subject does not notice any change in the quality of his visual experiences is perfectly explained by the fact that he does not perceive any difference between his current visual experiences and his memories.

Nonetheless, there is an empirical fact that challenges the idea that the destruction or replacement of the mechanisms that generate visual experiences brings with it the destruction of the visual memories of the subject. Cortical blindness is the total or partial elimination of vision generated by damage to the visual cortex. If the generation of memories concerning the qualitative character of visual experiences were so dependent on the visual cortex, we would expect that patients that exhibit cortical blindness reported the loss of some of their visual memories. However, this is not normally the case. Patients suffering from cortical blindness may report a loss of their visual abilities, but they simply do not say that they forgot the colour of a red apple, or the blue colour of the summer sky. It might be true that the visual cortex plays a certain role in the ability of recollecting past visual experiences, but it is clear that the elimination of this region of the brain does not necessarily brings with it the loss of this ability.

Second, the assumption that a system can only experience a change in his visual experiences through some recollection of previous visual experiences is not justified. It is

perfectly possible to compare a change concerning our visual experiences without comparing them with previous memories of these experiences. The version of the replacement argument that is presented by Zuboff in (1994) clearly illustrates this point. This version shares some elements of the one presented in the preceding section, but there is a difference whose importance will be clear at the moment of evaluating the objections against the replacement strategy. The version of the replacement thought experiment presented by Zuboff is also similar to the version he offers against the hypothesis of absent qualia. In particular, Zuboff adopts a coarser level of functional organization compared to the neural level adopted by the version that has been presented in this chapter. This level can be identified with the functional organization instantiated by the brain at the level of those parts of the brain associated to a determinate mental function.

As it is known, the visual cortex has two separate regions located in each of the two hemispheres of the brain. The left visual cortex processes visual signals that proceed from the right visual field, while the signals that have their origin in the left visual field are processed by the right visual cortex. Imagine now that during the replacement process the left side of Paul McCartney's visual cortex is removed and in its place is installed a gadget that has the same causal relationships that this side had with respect to the rest of Paul McCartney's brain. It is important to remember that the way the gadget works and produces electrochemical signals might be completely different from the way that the original side of the visual cortex works. Perhaps the easiest way for imagining the replacing gadget might be to conceiving it as composed by artificial neurons, in such a way that it is functionally equivalent, at the neural level, to the left side of Paul McCartney's visual cortex. This option may coincide with the way that the replacement process has been imagined along this chapter. This, however, is not necessary: although the gadget still needs to preserve all the relevant causal relationships that the original part had with respect to the rest of the brain and to produce the adequate electrochemical signals, it might do it in a completely different way. It may be, for instance, an artificial organ made out of plastic parts and silicon chips, and that has an internal “factory” for producing the adequate chemical neurotransmitters, which will be sent to those biological neurons connected to the gadget. It can also be imagined as a small space in the brain in which a number of homunculi perform

several calculations that allow them to generate signals that, after being transformed into appropriate chemical impulses, duplicate the way in which the left side of Paul McCartney's visual cortex works.

The key for arguing against the scenario depicted by the reductio hypothesis is that only the left side of Paul McCartney's brain was replaced by an artificial gadget, while the right organic side was left in its original place. If the inverted qualia hypothesis were true, Paul would find himself in a very odd situation: since his left visual cortex was not modified by the replacement process, all the experiences situated in his right visual field – controlled by the organic left side of his visual cortex – would be exactly as before: he will notice nothing strange or different. In contrast, the visual experiences in his left visual field, controlled now by the artificial right side of his visual cortex, will appear inverted with respect to his normal colour experiences.

Suppose that the scientists decided to stop the replacement operation at this point. They send Paul home and ask him to return the next day for resuming the process. When Paul leaves the hospital, he notices that the colour of most London buses appears red in his right visual field, that the colour of the sky is blue, and the foliage of most trees in the street is green. But the world would appear very different in his left visual field: the buses would appear green, the sky yellow and the leaves of an oak tree red.

> … our gadget replacement was of only the left half of the visual cortex. So if this replacement somehow resulted in an inversion or absence of qualia in the vision processed by the gadget, this inversion or absence on the right side of vision would clash with the necessarily unchanged qualia of visual memories and associations, as well as with the unchanged qualia of the other side of the visual field. Such a clash would make it absurd, in the now familiar way, that the pattern of mental functioning could not be reflecting a clash. So qualia inversion or absence cannot be what is happening in the gadget replacement. The experience must simply be the same.[63]

The argument can be presented as follows: if the qualitative character of the visual experiences of Paul McCartney were inverted after the replacement process, then only the

63 " Zuboff (1994, pp. 187-188)

experiences concerning the left side of the visual field would have been modified (since GV replaced only the right side of the Visual Cortex). But if the experiences in the left side of the visual field were modified, then there would be a clash with respect to the unchanged experiences of the right side of the visual field (since the replacement process did not touch the left Visual Cortex). Now, if there is a clash between the experiences of the left and right visual fields, then this has to be reflected in the behaviour, and in general, in the beliefs, desires and other mental states concerning these visual experiences. By assumption of the replacement scenario, the subject's behaviour would not have been different from the one he would have exhibited in case that his visual cortex had not been replaced by gadget GV, so his behaviour could not have reflected the clash. Therefore, the qualitative character of the visual experiences of Paul McCartney cannot have been inverted after the replacement process.

What is precisely the nature of the clash mentioned by Zuboff? If the replacement of the right side of his visual cortex had produced an inversion of his qualitative visual experiences, such inversion would have affected only the experiences of his left visual field. Now, by assumption of the replacement process, the left side of his visual cortex did not suffer any alteration, so the visual experiences in his right visual field – which are processed by the left visual cortex- are not modified either. If we assume that the replacement produced such inversion of qualia, then Paul McCartney would indeed be in a very odd situation: he would notice an obvious discordance concerning his visual experiences, and this would have left him extremely amazed.

But of course, we know that such change would not have affected Paul McCartney's behaviour. If we asked him “Hey, Paul, do you notice something odd with your sight?” he would answer negatively (if he wanted to give a sincere answer, of course). We know the reason: the gadget preserved the same causal relations to the rest of his brain that the original half of the visual cortex had, and thus, his behaviour could not have been modified by the replacement. This contradicts the idea that the clash between the visual experiences of Paul McCartney's left and right visual field had to be reflected in his behaviour, and thus, it is not possible that the gadget produced an inversion of his visual qualitative experiences.

Notice that there is a difference with the replacement process described in the last section. In the version of the thought experiment examined before, the visual experiences of Paul change when the switch is moved from position *a* to position *o* and vice versa (assuming, of course, that the inverted qualia hypothesis is true). In this case, the change in the visual experiences would be experienced *temporally*. Paul could not perceive that there is something odd with his visual perceptions at the same moment. However, in the case of the thought experiment proposed by Zuboff, the change in the visual experiences of Paul would be experienced *spatially*: if the hypothesis of inverted qualia were true, Paul would notice a contrast between the visual experiences of his right and his left visual field.

Conclusions

In the last section of this chapter, I presented three main reasons for showing that the objections presented by Van Heuveln *et al* against the dancing qualia version of the replacement thought experiment ultimately fail. Remember that according to the objection raised by Van Heuveln *et al*, it is not possible to be sure that there is one and the same individual before and after the switch is moved from position *o* to position *a*. It is perfectly possible to argue that the movement of the switch generates two different beings, each of them having its own phenomenal world. In such a case, it does not make sense to say that a single individual could have noticed a change in the qualitative character of the visual experiences, since the experiences before and after the replacement are owned by two different beings.

Against this claim, it was argued that the dancing qualia thought experiment presented by Chalmers involves a relatively small area of the brain. It is true that a replacement process like the one imagined by Van Heuvlen *et al* would not preserve the identity of the subject, but remember that this process involved the whole brain of the individual and not a small part. As we have seen, in order to assume that there is a noticeable difference between the qualitative experiences of two isomorphic beings, it is not necessary to conceive a difference that involves the replacement of more than ten per

cent of the brain. Thus, if the dancing qualia argument adopts this assumption, it is unlikely that the replacement of this part generates two different individuals.

Van Heuveln *et al* also argue that the process by which a subject recollects previous visual experiences involves the same mechanisms that generate these experiences. But if this were true, it would be possible to imagine that both the qualitative character of the current visual experiences of the subject and the qualitative character of his visual experiences depended on the physical nature of the visual cortex. So, it would be perfectly possible to imagine that the qualitative character of the visual experiences of Zero-Paul and Ten-Paul are different, and thus, that Zero-Paul would not notice any change when the switch is moved.

Against this claim, I have argued that if the visual memories of an individual were so dependent on the visual cortex, patients that exhibit cortical blindness would lose most of their visual memories. But as we have seen, this is not normally the case. Patients whose visual cortex is damaged may lose their visual abilities. However, they do not normally forget the qualitative character of their previous visual experiences.

Finally, I have argued against the assumption that a system can only experience a change in his current visual experiences by recollecting previous visual experiences. Remember that Zuboff conceives a replacement of only the right side of the visual cortex. If we assume that the replacement does not preserve exactly the same visual experiences than the ones generated by the organic visual cortex, then the subject will notice a contrast in his visual field. The left visual field will be as always, while the right visual field (which is controlled by the right visual cortex) will exhibit a colour inversion with respect to the other field. Nonetheless, and due to the fact that the gadget that replaces the right visual cortex is sending exactly the same signals to the rest of the brain, the behaviour of the subject, together with his beliefs, memories and other cognitive states, will not exhibit a change. Notice again that in this case the change of visual experiences would be experienced spatially and not temporally.

CHAPTER 5

The interchange version of the replacement thought experiment

Introduction

Remember again the objectives of the two different versions of the replacement thought experiment that were presented in the first chapters of this thesis.

> (NSS) There is a set of functional organizations F such that the capacity of a physical system P of generating conscious phenomenal experiences naturally strongly supervenes on the property of P of instantiating a member of set F.

> (NSS_2) There is a set F of functional organizations such that the qualitative character of the conscious phenomenal experiences of a physical system naturally strongly supervenes on the property of P of instantiating a member of set F.

Now, if thesis (NSS) is true, then the absent qualia hypothesis is naturally false. What the version of the replacement thought experiment presented in the second chapter of the thesis shows is the following: there is at least one functional organization (i.e., the functional organization of the brain of a conscious person at a neural level) such that its implementation by a physical system is a sufficient condition for generating conscious phenomenal experiences. Similarly, if thesis (NSS_2) is true, the inverted qualia hypothesis is naturally false. The version of the thought experiment presented in the third chapter of the thesis shows that the implementation of the functional organization of the brain of a conscious person at the neural level is a sufficient condition for the generation of the same qualitative experiences generated by the brain of this person.

The objective of the thought experiment that I will discuss and evaluate in this chapter, which is originally presented by Tye in (2006), is more ambitious. Its goal is to show that the absent qualia hypothesis is not only nomically (or empirically) false: according to Tye, it is conceptually impossible for a being who is a complete psycho-

functional isomorph of a normal human to lack phenomenal consciousness. As such, I will present it as supporting thesis (LSS):

> (LSS) There is a set of functional organizations F such that the capacity of a physical system P of generating conscious phenomenal experiences supervenes logically on the property of P of instantiating a member of set F.

Tye formulates the absent qualia hypothesis as posing an objection to functionalism, and more precisely, as an objection to a functionalist account of phenomenal consciousness. Tye states the absent qualia hypothesis as follows: "... it could be the case that a system that functionally duplicates the mental states of a normal human being has no phenomenal consciousness (no qualia)."[64] Tye suggests that the current orthodox position in philosophy of mind is to accept the claim that absent qualia are conceptually possible. In the actual world, the laws of nature simply do not allow mental differences between functional duplicates. However, there might be possible worlds where the laws of nature are different, and it might be the case that in some of them functional duplication does not suffice for mental duplication. Thus, from the fact that two individuals share the same functional organization does not logically follow that both of them share the capacity of experiencing conscious phenomenal states.

> Current orthodoxy in the philosophy of mind has it that absent qualia are at least conceptually possible. This is the view of all dualists about phenomenal consciousness and many materialists. I have come to think that orthodoxy is wrong. Proper and full a priori reflection upon the putative case of absent qualia demonstrates that they are impossible.[65]

An important characteristic of this version of the thought experiment that it is aimed to show that the claim that functional isomorphs do not share the capacity of having conscious phenomenal states is contradictory, and thus, that the negation of this claim is a necessary thesis. This contrasts with the version of the replacement thought experiment proposed in the second chapter of this thesis. The arguments derived from this version of

64 " Tye (2006, p. 140)

65 " Tye (2006, p. 140)

the thought experiment explicitly support the thesis that functional twins share the capacity of experiencing conscious phenomenal states in all worlds that share the same physical laws. In contrast, Tye thinks that functional duplication guarantees the presence of phenomenal consciousness in all logical possible worlds

> ... if absent qualia are conceptually impossible, then there cannot be a world that is physically just like the actual world in all respects and thus that contains creatures who are microphysical duplicates of normal human beings, where these creatures lack phenomenal consciousness. [66]

In the first section of this chapter I will define a being that is a psychofunctional duplicate of Paul McCartney, in the sense that the internal states of this being have the same causal roles that the internal states of Paul have with respect to other internal states and to inputs and outputs. The internal states of this being (which I call "A-Paul") will contrast with the internal states of Paul in that they are phenomenally inert. That a being like A-Paul is conceivable will constitute the *reductio* hypothesis adopted by this new variation of the replacement thought experiment.

In the second section of this chapter, I will present principle (P), as well as Tye´s assumption that a being that shares the same psychofucntional organization of a conscious being is capable of having beliefs, desires and other cognitive states. In this section, I discuss a thought experiment that shows that a system can behave in an intelligent fashion, but whose internal configuration indicates that it does not have intelligence at all. I argue, however, that the system described by Block is different from a functional duplicate of a conscious being. In contrast with the internal states of the system described by Block, a system like A-Paul exhibits the same causal relations with inputs, outputs and corresponding internal states of the internal states of Paul.

In the third section of this chapter, I will present the interchange version of the replacement thought experiment. Tye´s strategy begins by assuming, as a *reductio* hypothesis, that the thesis of absent qualia is conceivable. Thus, it would be possible to

66 " Tye (2006, p. 163)

conceive the existence of a system S' that is the functional duplicate of a conscious being S but whose internal states are phenomenally inert. Tye´s thought experiment consists in a partial interchange of the phenomenal states of being S with the non-phenomenal states of an isomorphic being S'. This interchange, however, will preserve the memories of each of these beings. The outcome of the process will be that the cognitive reaction to this interchange exhibited by S and S' will be different: S will think that he lost something valuable, namely, his conscious phenomenal states. In contrast, S' will think that he gained something of value. This difference concerning the cognitive reaction to the interchange contradicts a principle conceived by Tye as necessary:

> (P) Necessarily, if family F of mental states in being S has members that are one-to-one functionally isomorphic with the members of family F' of mental states in being S', where S and S' are themselves psycho-functional duplicates, then exchanging the two families preserves psycho-functional duplication

Now, if the cognitive reactions of S and S' are different, then they cannot be functionally isomorphic after the interchange of their respective phenomenal and non-phenomenal states. This, however, is inconsistent with principle (P). From this, Tye concludes that a being like S' is not conceivable, and thus, that the absent qualia hypothesis is logically false.

In the fourth section of this chapter, I will evaluate an objection presented originally by Van Gulick in (2012) whose objective is to show that the contradiction suggested by Tye does not arise. Van Gulick challenges principle (P) by arguing, first, that the existence of a being like A-Paul does not imply the possibility of a partial exchange between his non-phenomenal states and the phenomenal states of Paul. Second, Van Gulick argues that, even if we accept that such exchange is possible, it does not guarantee the preservation of the functional equivalence between Paul and A-Paul. If this is true, the *reductio* strategy of adopted by Tye is blocked: the resulting systems will exhibit a different cognitive reaction to the interchange process, but since principle (P) is not true, the contradiction suggested by Tye does not arise.

Finally, in the fifth section of this chapter, I present a strategy whose objective is to show that two beings that are psychofunctional duplicates, like Paul and A-Paul, can interchange their respective phenomenal and non-phenomenal states in spite of Van Gulick objections. However, I consider that this proposal ultimately fails. The general conclusion of this chapter is that the interchange version of the replacement thought experiment does not ultimately show that assuming the existence of a being like A-Paul leads to a contradiction. Thus, the thought experiment does not adequately show that the thesis of absent qualia is logically false, and it does not give an adequate support to thesis (LSS). It is important to notice that my claim is not that thesis (LSS) is false: there might be other ways of arguing in favour of the claim that phenomenal experiences logically supervene on the property of instantiating a determinate functional organization. My claim is that this version of the replacement thought experiment does not provide an adequate support for this thesis.

Section 1
Psychofunctional isomorphs

The argumentative strategy adopted by this new version of the replacement thought experiment will be better understood if we consider a case very similar to the one illustrated by Paul and his functional duplicate, F-Paul. Remember that, as it was defined before, Paul and F-Paul are almost physically identical, with the difference that, instead of an organic brain, F-Paul possesses an artificial brain that shares the functional organization of Paul´s organic brain at the neural level. Whether F-Paul had the capacity of having conscious phenomenal experiences or not was an open question. In this new case, however, Paul´s functional duplicate will be explicitly defined as lacking this capacity. In order to differentiate between these cases, this functional duplicate will be called A-Paul.

The functional isomorphism between Paul and A-Paul will be defined in a different fashion, which can be understood with the help of a strategy that was originally proposed by Lewis in (1972) and that has become a standard definition of functionalism. Briefly, the proposal says that a mental state *S* – a belief, a desire, or a conscious phenomenal state –

can be defined as occupying a certain causal role or function with respect to a number of other mental states, together with inputs in the form of sensory stimuli and outputs consisting in behavioural responses.

The method proposed by Lewis starts by enumerating a number of platitudes concerning the causal relations that exist among perceptual stimuli, mental states and behavioural responses. We may think that these platitudes take the form of a sentence like the following: when an individual is in a certain mental state and receiving such and such perceptual stimuli, this causes (with certain probability) this individual to go to such and such mental states and to produce such and such behavioural response. These platitudes may include causal generalizations concerning a mental state *S*. These generalizations establish the several relations that *S* has with other mental states, and with the inputs and outputs received and produced by the system to which *S* belongs. As an initial example, consider the following schematic functional definition of pain: pain tends to be produced by bodily damage, and tends to generate the belief that the body has been wounded, as well as a sensation of anxiety, fear and the desire of being out of that state, which in turn produces a behavioural response that can take the form of wincing or moaning.

Suppose that the mental state in question is the one that Paul is in when he smells a piece of Limburger cheese, and call it state S_{CH}. Among the several causal generalizations related to state S_{CH} are, for instance, that state S_{CH} tends to be produced by smelling a piece of Limburger cheese, that state S_{CH} tends to produce in Paul the memory of having eaten in a restaurant, that state S_{CH} tends to produce in Paul the desire of eating Limburger cheese, that state S_{CH} tends to produce in Paul a smile, and so on. Suppose now that there is a theory T composed by sentences expressing all these generalizations, which can be put in conjunction to form sentence T below (where each S_j denotes the names of other mental states, and each I_j, O_j specify the inputs and outputs received and produced, respectively, by the system to which S_{CH} belongs).

$$T\ (S_1 \ldots S_n, I_1 \ldots I_k, O_1 \ldots O_m)$$

The next step is to replace each state-name $S_1 \ldots S_n$ by a corresponding variable, and to bind each of these variables by an existential quantifier. The resulting sentence

(R_T) $\exists x_1 \ldots \exists x_n$ T ($x_1 \ldots x_n$, $I_1 \ldots I_k$, $O_1 \ldots O_m$)

is known as the Ramsey sentence of theory T. Sentence (R_T) offers the conditions under which Paul can be said to be in state S_{CH} (where x_{CH} is the variable replacing the name "S_{CH}"):

> Paul is in mental state S_{CH} if and only if $\exists x_1 \ldots \exists x_n$ [T ($x_1 \ldots x_n$, $I_1 \ldots I_k$, $O_1 \ldots O_m$) & Paul has mental state x_{CH}].

The definition of a mental state in terms of Ramsey sentences makes it easy to understand the idea that two mental states can be psycho-functional isomorphs. Two mental states are functionally equivalent in case they satisfy the same Ramsey sentence, that is, in case the relations they have with certain mental states, inputs and outputs are identical. Imagine, for instance, that for each mental state S_i had by Paul McCartney, there is a corresponding state S'_i had by A-Paul. In this case, the physical composition of A-Paul does not need to be specified: he can be made out of microscopical computers, tiny homunculi or almost any other material. To say that Paul McCartney and A-Paul are psycho-functional isomorphs, it is necessary that, for each mental state S_i satisfying a condition of the form

> Paul McCartney is in state S_i if and only if $\exists x_1 \ldots \exists x_n$ [T ($x_1 \ldots x_n$, $I_1 \ldots I_k$, $O_1 \ldots O_m$) & Paul McCartney has state x'_i].

(where x_i is the variable that replaces the name S_i) there is also a mental state S'_i that satisfies the condition

> A-Paul is in state S'_i if and only if $\exists x'_1 \ldots \exists x'_n$ [T ($x'_1 \ldots x'_n$, $I_1 \ldots I_k$, $O_1 \ldots O_m$) & A-Paul has state x'_i].

(where x'_i is the variable that replaces the name S'_i). Thus, mental states S_i and S'_i are psycho-functional isomorphs in case they have the same relations with other mental states, inputs and outputs. Also, two beings A and B (like Paul and A-Paul) are psycho-functional isomorphs if all mental states in A have a corresponding mental state in B that has the same relations with other mental states, inputs and outputs, and vice versa.

Section 2

Intelligent beings

The thought experiment that will be presented in this section describes a process in which the conscious phenomenal states of Paul are interchanged with the phenomenally sterile, but functionally equivalent states of A-Paul. The interchange process is conceived as preserving the phenomenal memories of Paul McCartney, as well as the non-phenomenal memories of A-Paul (if any). The strategy proposed by Tye is to show that this hypothesis leads to a contradiction that will arise as follows: after the interchange of internal states, the cognitive reactions exhibited by Paul McCartney and A-Paul will be different. One of them will think that he lost something that he greatly appreciated (his conscious phenomenal states) while the other, in contrast, will think that he won something valuable after the interchange (that is, the conscious phenomenal states of his functional twin). But if this cognitive reaction with respect to the outcome of the interchange process differs in such a way, then they cannot be functionally equivalent after this process. This, however, is inconsistent with a principle that, according to Tye, is both *a priori* and necessary.

> (P) Necessarily, if family F of mental states in being S has members that are one-to-one functionally isomorphic with the members of family F' of mental states in being S', where S and S' are themselves psycho-functional duplicates, then exchanging the two families preserves psycho-functional duplication.[67]

67" Tye (2006, p. 153)

Remember that Tye's objective is not only to show that the absent qualia hypothesis is in fact false. According to Tye, it is conceptually impossible that two functional duplicates – like Paul McCartney and A-Paul – differ in that one of them has conscious phenomenal experiences while the other does not. If the arguments that will be shown in the next paragraphs are sound, the absent qualia hypothesis is conceptually false: it involves a logical contradiction. In particular, the argument will show that assuming the existence of a psychofunctional duplicate that lacks conscious phenomenal states is logically inconsistent with principle (P).

Tye assumes the existence of a division between the cognitive and the phenomenal aspects of the mind, and suggest that the phenomenal side is the most difficult part of the mind-body problem. He suggests that the psychological aspect of the mind, that is, the aspect mostly concerned with cognitive states, like beliefs and desires, constitutes a problem that is indeed difficult, but not extremely puzzling for the philosophy of mind. According to his position, cognitive states do not present a difficult challenge for functionalism in particular, and suggest that a functionalist account of beliefs and desires should not be seen as particularly controversial.

> What seems to me clear is that any system that is a *complete* psycho-functional isomorph of me, that is, any system that duplicates my psychological states functionally across the board not only at the manifest, commonsense level but also at the level of science will be subject to beliefs as a matter of conceptual necessity.[68]

Tye assumes that, no matter their internal physicochemical composition, individuals that share the same psychofunctional organization of a person are also capable of having beliefs, desires and other cognitive states. Tye claims that, in contrast with conscious phenomenal states, cognitive states – beliefs in particular – are not especially problematic for functionalism. Tye argues as follows: the principles of psychological explanation seem to be easily applied also to individuals that share the same psychofunctional organization of a person. Tye suggests that the best way of explaining the behaviour of these duplicates will

68 " Tye (2006, p. 148)

be to attribute them beliefs and desires, just as we do when we want to explain the behaviour of a person. Imagine for instance that somebody asked Paul´s psychofunctional duplicate a question like the following one: "What was the name of the rock band you created with John Lennon, George Harrison, and Ringo Starr?" Assuming that he wanted to give a true answer, he would respond "The Beatles". If somebody asked him what was the name of the city where he met John Lennon for the first time, he would answer "Liverpool". If somebody asked him what is the name of the person in charge of the cinematography of the movie "Magical Mystery Tour", he perhaps would leave his seat, go to his bookshelf to consult the information required in the back side of a DVD box (I am assuming that he would not remember the name of that person) and once he finds the answer, he would say "Daniel Lacambre". Now, if we wanted to give an explanation of A-Paul´s behaviour, we would say that it is because he believes that there is a DVD box in the bookshelf where he can find the answer to that question, and that he has the desire of giving an accurate response. A-Paul´s behaviour allows the attribution of beliefs and desires, since it seems an adequate way for explaining his behaviour.

A problem concerning this proposal is that it is perfectly possible to imagine an individual that behaves in an intelligent way, but without being intelligent at all. Block (1981) offered a thought experiment for showing that the attribution of intelligence and of cognitive states purely from behavioural grounds is mistaken. Two systems can be behaviourally alike (concerning their actual and potential behaviour, their behavioural dispositions and their counterfactual behavioural properties) and nonetheless differ in the way that they process information that mediates between perceptual inputs and behavioural outputs in such a way that this difference will determine that, while one of them is fully intelligent and has cognitive states, the other is not. The nature of the systems internal processes is essential for determining whether they have intelligence or not.

Block's strategy is to describe a machine that is able to pass a behaviouristic test of intelligence, but whose internal configuration shows that it does not have intelligence at all. The logical possibility of this machine will show that behavioural criteria do not suffice for attributing intelligence, or the presence of cognitive states like beliefs or desires. The test

considered is the famous Turing test, which was originally proposed as a way for determining if machines can think, and that is commonly interpreted as proposing sufficient conditions for intelligence. Turing considered that, due to the difficulty to disambiguate the terms occurring in the question 'can machines think?' it is better to consider a different strategy for dealing with this problem, although formulated in a comparatively unambiguous way. Briefly, the Turing test consists in playing a game known as the 'imitation game' which is played as follows: three players, a person, a machine and an interrogator or judge start a conversation. Both the person and the machine try to appear human, and if the interrogator cannot distinguish between them after a certain period of time, it is said that the machine passed the test.

Suppose the duration of the test is one hour and that the language used is English. Let us define the set *T* of sequences of English sentences that can be typed in one hour as the set of *typable* sequences of sentences, and the proper subset *S* of *T* the set of all *sensible* sequences of sentences, defined as those sequences that are naturally interpreted as conversations. The machine's memory contains the set of all sensible sequences of sentences in the form of a list, and when the interrogator emits sentence *A*, the machine searches on the list a sequence that starts with sentence *A* and answers the interrogator by emitting the next sentence *B* on the sequence. When the interrogator types the next sentence, the machine performs the same procedure, and so on, until the test ends. The machine proposed by Block is programmed in a way that allows it to produce a sensible verbal output for all possible verbal inputs produced by the interrogator, so it won't be possible for him to distinguish the difference between this machine and a genuine intelligent being. The idea behind this experiment is that a device that has the intelligence of a toaster (or perhaps better, the intelligence of a vending machine) will be able to pass a behavioural test of intelligence. The thesis that intelligence can be defined in purely behavioristic terms is, therefore, threatened by the logical possibility of this supposedly stupid machine. If this unintelligent machine is a genuine logical possibility, then it is possible for an unintelligent entity to pass the Turing test, and since this test is designed from a behavioristic point of view, Block claims that behaviorism is false.

But although Block´s thought experiment might undermine behaviourism, there is a key difference between the system described by Block and Paul´s functional duplicate. The goal of the thought experiment presented by Block was not only to show that behaviourism is false, but also that they way in which an entity processes information is essential for determining whether it has cognitive states or not. This is precisely the difference between Block´s machine and Paul´s functional duplicate. In the former case, the internal states of the machine described by Block are completely different from the internal states of Paul. In the latter case, the internal states of A-Paul have been defined as being isomorphic to those of Paul.

> ... it seems that our concept of a belief or a desire is the concept of a state that plays an appropriate functional role. So, it seems that my functional duplicate, like me, has beliefs and desires. And these beliefs and desires explain his behavior. The principles of psychological explanation apply to him just as they do to me. [69]

Thus, it is true that a being like the one described by Block can behave as an intelligent being without having beliefs, desires or other cognitive states. However, in the case of a psychofunctional duplicate of a conscious being, like A-Paul, there are internal states that function in the same way as the beliefs and desires of Paul, that is, they have the same relations with respect to corresponding internal states and to inputs and outputs.

Section 3

The interchange strategy

Let's consider now the interchange strategy in favour of the claim that the absent qualia hypothesis is conceptually false. Suppose that it is possible to conceive a psycho-functional duplicate of Paul McCartney that differs from him in that his internal states, in spite of being psychofunctionally identical to those possessed by the original Paul, lack phenomenal properties. To mention an example, imagine that Paul McCartney is in a certain phenomenal conscious state S_E when he hears the lower string of his bass emitting

69 " Tye (2006, pp. 147-148)

the sound E_1 (assuming that his bass is standardly tuned). Now, since A-Paul is a psycho-functional duplicate of Paul McCartney, he will possess a corresponding internal state S_E' that has the same relations that S_E has with respect to other mental states, inputs and outputs. The difference will be that, while S_E is a mental state that has a determinate phenomenal character, state S_E' is phenomenally sterile. The hypothesis that it is possible to conceive a psycho-functional duplicate like the one exemplified by A-Paul will constitute the *reductio* hypothesis in the argument offered by Tye.

Now, the way in which the notion of psychofunctional duplicate has been understood suggests the possibility of an interchange of families of functionally equivalent internal states. For instance, we may imagine that the interchange deals with those internal states that, in Paul McCartney, concern auditory perceptions. Since A-Paul is psychofunctionally identical to Paul McCarntey, A-Paul has internal states that are functionally isomorphic (in the sense explained earlier) to these auditory states, only that, according to the *reductio* assumption, they are phenomenally inert. Tye also mentions that the interchange does not need to be restricted to families of mental states: the interchange can also be of individual mental states M and M'. For instance, the state S_E in which Paul McCartney is when he hears the vibration of the lower string of his bass will have a corresponding state S_E' that is functionally identical, but that lacks its particular phenomenal properties.

Imagine now that a group of scientists has built a very complicated machine similar to the one described earlier. In this case, however, this machine is capable of interchanging the phenomenal states of Paul McCartney and the non-phenomenal states of his psychofunctional isomorph, A-Paul. Briefly, the machine works as follows: when the individuals are introduced in the machine their heads are put into a pair of helmets. When the machine is turned on, a huge number of tiny robots are introduced into the heads of Paul McCartney and A-Paul. These robots then perform a number of internal changes inside their heads and, when the process is completed, Paul McCartney and A-Paul will have interchanged phenomenal and non-phenomenal internal states. Paul has lost all his conscious phenomenal states, which were replaced with the non-phenomenal states of A-

Paul, and at the same time Paul has preserved his previous phenomenal memories. Similarly, A-Paul has lost all his non-phenomenal internal states, which were replaced by the phenomenal states that originally belonged to Paul McCartney. Also, the non-phenomenal memories of A-Paul have been preserved. This is the way in which Tye conceives the outcome of the interchange process:

> The result of these changes is that there is a partial exchange of phenomenal states and nonphenomenal states between the two people, [...] This exchange is such that were I and my functional duplicate [...] to agree to undergo the exchanger operation (as it comes to be called), I would lose all my phenomenal states, *other than those that are phenomenal memories,* and I would have them replaced by corresponding ersatz phenomenal states. *NN* would lose all his ersatz phenomenal states, *except those that are ersatz phenomenal memories,* and he would have them replaced by corresponding phenomenal states.[70]

Before continuing, it is important to note that the scenario just described does not specify the nature of the physical realizers of the phenomenal and non-phenomenal states possessed, respectively, by Paul and his psychofunctional isomorph, A-Paul. In consequence, the description of the scenario does not include an explanation of what these miniature robots physically do in order to produce the interchange depicted in the thought experiment. Now, why might the specification of the physical realizers of these states be important? According to the way this strategy is originally described by Tye, the interchange process is conceived at the level of mental states, and not at the lower level of their physical realizers. Whatever the physical realizer of a certain mental state might be, or of a family of mental states, is considered relevant as long as it brings an interchange at this higher level. In some cases, the precise way in which these physical realizers are modified in order to generate the interchange at the level of phenomenal states is not difficult to conceive. Remember, for instance, the initial version of the replacement thought experiment described in the introduction of the thesis: suppose that the subjects involved in the interchange operation just described are Paul and F-Paul. In this case, the *reductio* assumption might be that F-Paul – who has an artificial brain that implements the functional organization of Paul McCartney at a neural level – lacks conscious phenomenal

70 " Tye (2006, p. 155)

experiences. In such a case, the interchange will be between the phenomenal states of Paul and the nonphenomenal states of F-Paul. In this case, however, we know what the physical realizers of these states are: organic neurons in Paul´s case, and miniature computers in the case of F-Paul.

The interchange process, however, is not clear when the psychofunctional isomorph of Paul has different physical realizers. Consider, for instance, Searle´s brain simulator, or Block´s Chinese nation. In the next section of this chapter we will see an objection that concerns the possible physical realizers of the internal states of Paul McCartney and A-Paul. This objection does not question the possibility of psychofunctional duplicates, but identifies a problem concerning the idea that these psychofunctional duplicates can interchange internal states, preserving at the same time their supposed psychofunctional isomorphism. Meanwhile, assume that the interchange described by the thought experiment is feasible, and lets see how can it be argued that to assume the absent qualia hypothesis leads to a contradiction.

The next step is to elucidate the nature of the cognitive responses exhibited by Paul and A-Paul to the interchange operation. Initially, we may imagine that Paul McCartney, after the process has been completed, will become amazed at the outcome. He will notice that he has lost something very valuable, namely his phenomenal conscious states. He would realize, for instance, that he could not perceive the taste of a pint of ale, the characteristic sound of his bass, the red colour of London buses, or the pain caused by smashing his toe with a hammer. Note that, by assumption of the thought experiment, the process preserved all the phenomenal memories that Paul McCartney had before the interchange operation. This allows him to notice the difference between his current psychological situation and his past phenomenal experiences.

The situation of A-Paul after the interchange operation exhibits a great contrast with respect to Paul´s: he now has conscious phenomenal experiences. He can now taste the flavour of a pint of ale, hear the sound of an orchestra, perceive the smell of a rotten egg and feel the painful sensation of smashing his fingers with a hammer. Analogously with the

previous case, A-Paul´s non-phenomenal memories were preserved after the interchange operation, and these memories allow him to compare his present situation with the one before the interchange operation.

Now, imagine now that a group of doctors propose Paul McCartney the operation just described. The doctors tell him that the operation would be completely safe. Moreover, he will be assured that the operation will preserve all of his cognitive capacities, together with his past phenomenal memories. The only drawback, however, will be that the interchange operation could not be reversed. Given this information, would Paul accept the operation? Note that he has sufficient information concerning the outcome of the operation: he is aware that his cognitive capacities will not be diminished, and further, that his memories concerning his past phenomenal experiences will still be there after the operation. He has a sound basis for deciding whether to accept the interchange operation or not. Now, assuming that Paul McCartney enjoys the taste of a pint of ale, and the sound of his bass, the answer would be most likely negative. Of course, we can imagine a situation in which Paul has a terrible disease, and he is informed that, unless he accepts a medical treatment involving extremely painful procedures, he could not be saved from death. In this case, Paul might happily accept to be operated. Perhaps the option of living without conscious phenomenal experiences is much better than living in constant pain. But if we assume that Paul is completely healthy, and that he appreciates (at least most of the time) his conscious phenomenal experiences, it seems clear that he will reject the operation. Again, losing his conscious states means losing something that he regards as valuable.

But let´s consider again principle (P). If this principle is true and we assume that the interchange operation just described is conceivable (i.e., if we assume that the interchange between the phenomenal and non-phenomenal states of two isomorphic beings is conceivable), then Paul and A-Paul must be psycho-functional duplicates after the operation. This is precisely what principle (P) assured: exchanging one-to-one isomorphic families of mental states of mutually psycho-functional duplicates must preserve psycho-functional duplication. But this evidently clashes with the earlier discussion concerning the cognitive responses of Paul and A-Paul to the interchange operation. As we have seen, Paul

´s evaluation of the outcome of the interchange operation will be mostly negative: after comparing his current situation with his earlier phenomenal memories he will notice that he lost the capacity of perceiving colours, of tasting a good slice of Wellington steak, or of hearing the sound of a musical instrument. A-Paul´s evaluation, in contrast, will be extremely positive: based on his non-phenomenal memories, he will thought of his new situation as being richer and much more exciting than his previous mental life: after all, he is now able to perceive colours, of tasting a nice steak, or of hearing the music of the Beatles.

The contradiction that dismisses the reduction assumption – namely, that the internal states of the psycho-functional isomorph of Paul do not have any qualitative properties – is now obvious. The preceding reasoning shows that the cognitive reactions of Paul and A-Paul to the interchange operation are different: on the basis of his phenomenal memories, Paul will believe that he lost something valuable (his conscious phenomenal experiences). Also, on the basis on his non-phenomenal memories, A-Paul will believe that he gained something of value. But according to principle (P) (that Tye regards as a necessary, conceptual truth) Paul and A-Paul will still be psychofunctionally duplicates after the interchange operation. Thus, their corresponding cognitive reactions to the interchange cannot be different. This is precisely the contradiction detected by Tye. But if the assumption that the internal states of A-Paul do not have any qualitative properties implies a logical contradiction, it must be logically false. This shows the absent qualia hypothesis – the *reductio* assumption of the argument, which is understood as the claim that psycho-functional isomorphs of human beings lack qualitative states – is also logically false.

> ... if [Paul and A-Paul] are genuine functional duplicates, then there *cannot* be a difference in our cognitive reactions of the sort I have been insisting upon (a difference that will manifest itself in a difference in verbal behavior, for example). But there must be such a difference, I have argued. That is what the above reasoning compels us to conclude. The contradiction reached here shows that it is *not* conceptually possible for me to have a complete functional isomorph who undergoes merely ersatz phenomenal states. Necessarily, any system that functionally duplicates me is phenomenally conscious. The absent qualia hypothesis, therefore, is false even on its weakest interpretation.[71]

71 " Tye (2006, p. 159)

A possible way of arguing against this outcome is to say that, although A-Paul gained something valuable after the interchange operation, his cognitive response will be identical to Paul´s: he will believe (falsely) that he lost something valuable. The problem with this approach is that A-Paul´s belief is generated introspectively in virtue of his previous memories. But since his power of introspection and his memories are working as well as Paul´s, it is simply not clear how A-Paul could have reached the belief that he lost something of value. This first approach seems to be clearly inadequate. Perhaps we could accept that Paul has lost something valuable, while A-Paul has won it, but there will be no difference in their cognitive reactions. By comparing his actual phenomenal experiences with those in his memory, Paul would think that he lost something important. However, A-Paul´s reaction with respect to the outcome of the operation will be identical: he will sincerely believe that something valuable has disappeared. The difference, of course, is that F-Paul´s belief is false. The problem, of course, is that the way in which F-Paul´s false belief is generated is left unexplained. Remember that one of the assumptions of the thought experiment is that F-Paul, as a functional isomorph of Paul McCartney, will have cognitive capacities like introspection. He is able to examine his current mental states and to compare them with his previous memories.

Section 4
Objections to the interchange strategy

In (2012) Van Gulick presents an objection against Tye's thought experiment. Van Gulick's strategy is to argue that the supposed contradiction identified by Tye simply does not arise, and consequently, that the *reductio* is blocked. As we have seen, a crucial premise of Tye´s argument is principle (P), according to which the interchange of one-to-one isomorphic families of internal states between two psycho-functional duplicates preserves psycho-functional duplication. Van Gulick identifies two crucial questions related to principle (P):

> (a) Would the conceivability of A-Paul entail that a partial exchange of corresponding states between Paul and A-Paul was also conceivable?

(b) Must this exchange leave Paul and A-Paul functionally equivalent to each other, as principle (P) asserts?

To the first question, Van Gulick answers as follows: even if it were possible to conceive a system like A-Paul (that is, a qualia-lacking system that is a psycho-functional isomorph of a conscious being) it is extremely doubtful that the corresponding phenomenal and non-phenomenal internal states of these systems can be interchanged. To the second question, Van Gulick´s response is clearly negative: even assuming that the proposed interchange is possible, we cannot be sure that we will end with two functionally identical systems. But in either case, the contradiction suggested by Tye does not arise. In the first case, the supposed interchange of phenomenal and non-phenomenal states between Paul and A-Paul would not be granted, and thus, Tye´s argument would lack an essential premise. In the second case, and even assuming the possibility of such an interchange, the cognitive responses of Paul and A-Paul to the operation will be different, but since they cannot remain psycho-functionally isomorphic after the interchange of phenomenal and non-phenomenal states, there will be no inconsistency with principle (P), and thus, Tye´s argument would be blocked.

First objection: it is not obvious that the internal states of two isomorphic systems can be interchanged.

Lets begin with the first questionable assumption identified by Van Gulick, namely, that an interchange of phenomenal and non-phenomenal subsets of internal states between two psycho-functionally isomorphic beings is conceivable. Van Gulick reminds us first that the question whether the interchange is empirically feasible or not is not relevant in the context of Tye´s argumentation: what matters is the logical possibility of such an interchange. But even in this case, and when the particular details concerning how the interchange can be realized are considered, the conditions of this realization are not obvious.

Remember that the argument presented by Tye is designed as a response to the cases illustrated by thought experiments like Block´s Chinese Nation and Searle´s system of water pipes already discussed in the first chapter of this thesis. As it has been mentioned before, a way for setting up the hypothesis of absent qualia has been to suggest the possibility of extremely unconventional systems that realize the functional organization of conscious beings. The claim that these systems lack conscious phenomenal properties is the target of Tye´s argument. Now, it seems to be clear that an interchange of internal states like the one depicted in the thought experiment proposed by Tye must be conceivable also in these cases. Otherwise, the result of this thought experiment cannot be fully general. But – Van Gulick argues – if Tye's goal is to argue that functional duplicates that do not share the same phenomenal conscious states are conceptually impossible, it is not sufficient to exhibit that a contradiction follows from imagining a single case. The argument must cover all these unconventional realizations. But if Van Gulick is right, the possibility of interchanging a subset of functionally equivalent internal states between a being with a brain and one of its isomorphic bizarre realizations is far from being obviously conceivable.

Remember again Block´s Chinese Nation thought experiment. A network of people equipped with radios send and receive data in such a way that the whole process is isomorphic to the pattern of signal interchanging in the brain of a conscious person. The role fulfilled by these people will depend, among other things, on the level of functional organization of the network. Perhaps a number of people inside the Chinese Nation form a structure that corresponds to the Visual Cortex of that conscious person. Or perhaps the functional organization of the Chinese Nation is at the neural level, and in such a case, the structures formed by people correspond to each of the neurons of the brain. In any case, the Chinese Nation can be said to implement the functional organization of the brain at a certain level when the whole system is divided into an appropriate number of parts with a corresponding physical state, in such a way that the causal dependency relations among these parts and the inputs and the outputs received by the system mirrors the abstract specification of the functional organization of the brain at the selected level. How can we conceive an interchange of the physical realizers of corresponding internal states between the Chinese Nation and this conscious person? Van Gulick notices that the property of a

certain structure of realizing a certain role inside a system will depend on the causal context of the larger system within which this structure is embedded. In the case of the Chinese Nation and the conscious person, an interchange of phenomenal and non-phenomenal internal states between them will involve an interchange of the physical realizers of these states. In the case of conscious states related to visual experiences, for instance, this might involve the interchange of structures associated to the Visual Cortex. The problem is that it is simply not obvious how a structure formed by people interchanging radio signals can make a relevant causal contact with structures formed by organic neurons that work by interchanging electrochemical signals.

> Any such exchange of token states would require more than the mere physical interchange in space of the two sets of realizers. At a minimum the exchange of mental states would have to involve the respective sets of realizers making relevant causal contact within their newly surrounding systemic contexts, and it is not obvious that doing so would be possible in all such cases. How would neural states be interchanged with nation-of-China states without destroying the respective systems and thus any possibility of causal engagement?[72]

But according to Van Gulick, in order for achieving Tye's aim – that is, in order to show the impossibility of conceiving isomorphic systems that do not share the same conscious phenomenal states – it is necessary to show that the physical realizers of this particular mental ability are able to maintain the same causal relations with the new systemic contexts, no matter the particular characteristics of these physical realizers. But that this is possible in the case of Paul´s neural states and the Nation-of-China states is far from being obvious.

> Given the enormous diversity in how functional organizations might be realized, not only in their particular concrete structures, but also in how those structures must interact at multiple levels to produce the requisite total organization, it is not at all certain that interchanges between isomorphic sets of states are always conceivable between systems that are equivalent relative to a given psycho-functional specification [...] Thus contrary to Tye's claim, the conceivability of absent qualia functional duplicates does not entail the conceivability of a partial exchange of isomorphic states of the sort he

72 " Van Gulick (2012, pp. 279-280)

proposes between conscious beings and their duplicates. The conceivability of such duplicates does not in itself guarantee the conceivability of partially exchanging states with them.[73]

Second objection: even assuming that an interchange of internal states between two isomorphic systems is possible, this does not mean that the interchange preserves the functional isomorphism between them.

The second problem identified by Van Gulick seems to be even harder. Suppose for a moment that the interchange described by Tye is effectively conceivable: it is possible to interchange different physical realizers of functionally equivalent internal states in isomorphic systems. However, eve if Tye is right in this, it does not follow that this interchange preserves psycho-functional duplication. The problem arises from the way in which an internal structure plays a role inside the system in which it is embedded. That a certain structure is capable of playing a certain role depends, first, on its particular causal profile: on the physical effects of this structure inside the system to which it belongs. Second, it also depends on the causal profiles of the other structures that belong to the system. An organic neuron fulfils a determinate role inside the brain insofar as it has a certain causal effect on other neurons, a causal effect that is determined by its capability of receiving, processing and sending electrochemical signals. A certain structure composed by people inside Block´s Chinese Nation has also a certain causal effect on other structures due to its capacity of receiving, sending and processing radio signals to other structures inside the system to which it belongs.

Imagine now that certain instance of Block´s Chinese Nation is psycho-functionally equivalent to Paul McCartney. This instance of the Chinese Nation has a family of non-phenomenal states F_{CH} that are isomorphic to family F_P of phenomenal states of Paul. In the case of the Chinese Nation, the physical realization of family F_{CH} of non-phenomenal states involves structures of Chinese people, while in the case of Paul the physical realization of family F_P of phenomenal states involves neural structures. These structures have certain

73" Van Gulick, (2012, p. 280)

causal effect inside their corresponding systems due to the particular way in which they interact with the other structures of their respective systems. But if the structures that realize families F_{CH} and F_P of isomorphic states are interchanged, it is simply not clear how they can work in the new systems in such a way that the functional equivalence between them is preserved. How can the structures that work through radio signals interact with a system whose basic structures work through electrochemical signals? Van Gulick illustrates problem with two systems F1 and F2 whose physical realizers are, respectively, gears and hydraulic mechanisms:

> The fact that the gears of F1 and the hydraulics of F2 enable them to function equivalently within their original causal contexts in no way implies that they would engage their reversed contexts in equivalent ways. Indeed they might fail to engage in any useful way at all, and even if they did engage, the results might be quite disparate in the two cases. The gears of F1 would not likely interact with the hydraulics of S2 in a way that mirrored that between the hydraulics of F2 and the gears of S1. The particular token states of F1 and F2 might continue to function within their isolated families but interact with their larger systemic surroundings in very dissimilar ways.[74]

Against Tye´s claim, Van Gulick concludes that principle (P) is not a necessary truth. First, that we can conceive a being like A-Paul – that is, a being that is psycho-functional isomorph to Paul – does not mean that we can conceive an interchange of isomorphic mental states between them. The physical realizes of these mental states have certain causal profile inside their original systems that might not match with the new systems after the interchange. Second, even if we accept the possibility of such interchange, it is not granted that Paul and A-Paul preserve their mutual psycho-functional isomorphism.

74 " Van Gulick, (2012, p. 281)

Section 5

A possible response

Remember again the two reasons offered by Van Gulick for claiming that the contradiction in Tye's reductio argument is blocked: the first is that, even if a psycho-functional isomorphism between a human being and another physical system is conceivable, it does not follow that their corresponding psycho-functionally states can be interchanged. The second is that, even it is assumed the conceivability of such interchange, it does not follow that the systems remain psycho-functional isomorphic after the interchange. By arguing in favour of these affirmations, Van Gulick intends to show that there is no contradiction, and thus, he concludes that Tye's argument is blocked.

It is clear that in some cases, the physical realizers of families of internal states of two isomorphic systems cannot be interchanged in such a way that they causally interact with their new surroundings. When the physical nature of the realizers of corresponding families of isomorphic internal states is unknown, it is not possible to determine the conditions for interchanging these families of internal states. Due to the wide physical variety of structures that can realize these internal states, we cannot assure that the interchange of these structures brings with it an interchange of respective isomorphic families of internal states. But in other situations the problem is simply not present. A case that clearly illustrates this situation is F-Paul´s artificial brain, which implements the same organization of Paul´s organic brain at a neural level. Remember that for each organic neuron inside Paul´s brain, there is a corresponding artificial neuron inside F-Paul´s artificial brain that performs the same input-output function, in the sense that when it receives a certain electrochemical signal, the artificial neuron process it and then sends another electrochemical signal to the rest of the brain.

Assume now that certain organic structures inside the brain of Paul and certain artificial structures inside the brain of F-Paul physically realize, respectively, families F and Fəof internal states, which are isomorphic in the sense described earlier. The difference, again, is that family F has phenomenal properties, while family Fəlacks them. But note that

in this case the causal profile of organic and artificial neurons is identical: an organic neuron receives as its input a determinate electrochemical signal, processes it and generates an electrochemical output that is in turn sent to other organic neurons inside Pauls brain. An artificial neuron can be defined as doing exactly the same. Of course, the way in which this signal is processed does not need to be identical to the way in which the organic neuron does it: this artificial neuron might be simply a capsule where a little homunculus generates the adequate signal by following a set of instructions. Note that in this case, the physical realizers of families of internal states F and Fɘare clearly interchangeable. The reason is that they share the same causal profile: both organic and artificial neurons send the same electrochemical signals to the rest of the brain.

With this in mind, we can argue as before: we adopt, as our reductio hypothesis, that F-Paul´s internal states lack phenomenal properties, in spite of being functionally isomorphic to Paul´s. Paul and F-Paul are then put in the machine described earlier, where tiny robots are introduced into their heads and make the necessary changes in their brains. In this case, however, we certainly know what these little robots are doing: they "disconnect" the neurons inside Paul´s brain that form the structures that realize family F of internal states, and install them in the places formerly occupied by those artificial structures that realized family Fɘof internal states inside the artificial brain of F-Paul. Equally, these artificial structures are in turn installed in the proper places of Paul´s brain. The argument can now proceed as before: after the interchange, Paul would think that he lost an important part of his mental life, namely, the conscious phenomenal experiences associated to the family of states F, and he would consider the outcome of the operation as negative. Meanwhile, F-Paul would think that he gained something valuable: the conscious phenomenal experiences lost by Paul. F-Paul would consider this a positive result, since this means that his mental life is richer than before. This difference in the cognitive reactions of Paul and F-Paul contradicts principle (P): if they remain psycho-functionally isomorphic after the replacement, then there cannot be a difference in their cognitive reactions to the operation. Thus, the assumption that F-Paul´s internal states are phenomenally inert must be false.

Of course, the point of Van Gulick is that this strategy is not general enough. The interchange proposed by Tye can be conceived between beings whose internal states have physical realizers that, once interchanged, can interact causally without a problem, like in the case of Paul and F-Paul. But what would happen with the bizarre systems proposed by Block (the China-body system) and the brain simulator made from water pipes proposed by Searle? As Van Gulick suggests, if Tye intends that his thought experiment achieves a certain level of generality, it does not seem sufficient to imagine the thought experiments with two beings whose physical realizers can be obviously interchanged.

There is, however, a strategy I want to propose for arguing in favour of the claim that, even in the case of these bizarre systems, families of mental states can be interchanged. Ultimately, I think that this strategy will not succeed, but not exactly for the reasons suggested by Van Gulick. As I will argue in the General Conclusions, the most troubling issue concerns the physical nature of the signals interchanged by these isomorphic systems. Meanwhile, the strategy I suggest in the paragraphs below will serve as a motivation for the discussion presented in the General Conclusions of the thesis.

Suppose that the brain of Paul implements, at a neural level, a certain functional organization O that can be abstracted into a Combinatorial State Automaton CSA_O. For each neuron inside Paul´s brain, there is a corresponding vector V of CSA_O that determines the precise input-output function of this neuron (call it neuron n_0). Now, imagine that the government of China – for using the example proposed by Block –has decided to build a system made from Chinese people that is a psychofunctional duplicate of Paul. For each neuron inside Paul´s brain, there will be a structure made from Chinese people that implement the same vector. In particular, there will be a structure CH that implements vector V. Now, it is clear that the formalization of this vector does not include the nature of the input and output signals. In the case of the neurons inside Paul´s brain, these input and output signals have an electrochemical character. In the case of the Chinese Nation, these signals might be verbal and are transmitted, perhaps, by radio.

If the brain of Paul and the Chinese Nation system implement the same functional organization by implementing the same CSA, there will be structures made of Chinese people that perform the same functional role of neural structures inside Paul´s brain. Assume, for instance, that there is a neural structure N that realizes a certain family F of phenomenal internal states inside Paul´s organic brain, while the Chinese Nation structure CH implements family Fə of non-phenomenal internal states inside the Chinese Nation system. It is clear that structures N and CH realize families F and Fəof internal states in virtue of their respective causal profiles. This is precisely the reason that prevented the interchange proposed by Tye. But imagine now the following scenario: imagine that we connect to structure CH a transmitter device T that works as follows: T receives the verbal signals that are produced by structure CH and transforms them into radio signals. Imagine also that we remove the neural structure N that physically realizes family F of internal states inside Paul´s brain, and in its place we install a receiver device R whose function is, first, to receive the radio signals generated by device T and transforms them into electrochemical signals that are in turn sent to the rest of Paul´s brain. The process can be illustrated schematically:

Structure CH (Verbal signals) ⇒ Transmitter T (Radio signals) ⇒ Receiver R (Electrochemical signals)

The same procedure can be applied to the Chinese Nation system: to begin with, we know that there is a structure N inside Paul´s brain that realizes family F of internal states. We can connect to this structure a similar transmitter device Tə(of course, with a much smaller size than the former) that, after receiving the electrochemical signals generated by structure N, transforms them into radio signals. We now remove structure CH inside the Chinese Nation system and install in its place a device Rəthat receives the radio signals generated by Təand transforms them into verbal signals that can now be sent to the rest of the Chinese Nation system.

Structure N (Electrochemical signals) ⇒ Transmitter Tə(Radio signals) ⇒ Receiver Rə (Verbal signals)

Now, the first problem identified by Van Gulick was that an interchange like the one proposed by Tye could not be accomplished due to the size of the corresponding realizers of families of internal states: evidently, we could not replace a neural structure with a structure of the Chinese Nation system without destroying Paul´s brain. But in the procedure just described, devices T and R are designed to fit physically into the corresponding systems. The second problem was that the causal profiles that allow structures N and CH to realize families F and Fɔof isomorphic internal states are different, and thus, an interchange of these structures would not guarantee an adequate causal interaction with the rest of the system. However, devices E and R have been described as transforming the signals from one system into signals that fit the causal profile of the other.

The crucial issue, however, is whether we can understand this process as an interchange of families F and Fɔof internal states between Paul and the Chinese Nation system. It can be suggested that, since the respective physical realizers N and CH of these internal states are not physically interchanged, but remain in their corresponding systems, there is not a real interchange of families F and Fɔof internal states. However, I think that we have good reasons to affirm that the procedure just described can be understood as interchanging these families of isomorphic internal states between Paul and the Chinese Nation system. Note that when device R is installed into Paul´s brain in place of neural structure N, there is a causal chain of signals that begins when structure CH emits a certain verbal signal. Clearly, device R generates certain electrochemical impulses in virtue of the instructions it receives, via radio, from device T. Device T, in turn, generates this radio emission by transforming verbal signals received from structure CH. If by some mistake structure CH sends a different verbal signal, then device T would transmit a different radio signal, which after being received by device R, would in turn generate a different electrochemical signal. The same happens when device R is installed into the Chinese Nation system in place of structure CH, although in this case, it generates a determinate verbal signal in virtue of the radio instructions sent by device T, which in turn are generated in virtue of the particular electrochemical signal produced by neural structure N. The role

of device R is simply to receive radio signals and to transform them into signals that have a causal effect into the respective systems.

Conclusions

In spite of these considerations, I think that the strategy presented above does not ultimately show that there is a genuine interchange between the phenomenal and non-phenomenal states of Paul and the Chinese Nation system. According to the strategy presented in this section, we can interchange the phenomenal and non-phenomenal states of isomorphic subjects by attaching transmitters and receivers to the physical realizes of these states. In order to assure a causal match with the new surrounding systems, it was stipulated that the receivers transform the radio signals into electrochemical or verbal signals capable of interacting, respectively, with Paul´s brain and with the Chinese Nation system. However, it can be argued that the respective phenomenal and non-phenomenal character of the states of these isomorphic systems are generated not because they fulfil a certain functional role inside their respective systems, but simply because the signals interchanged by their corresponding physical realizers have a determinate physical character. It can be argued that the physical realizers of the states of Paul are capable of generating phenomenal experiences in virtue of the transmission of electrochemical signals. Conversely, the non-phenomenal character of the states of the Chinese Nation system depends on the transmission, by their physical realizers, of verbal signals.

Now, remember that the neural version of the replacement thought experiment faced a similar problem. If it is stipulated that the neural replacements transmit the same electrochemical signals interchanged by the organic neurons of the brain of a conscious person, a system composed of artificial neurons that shares the same functional organization of this brain may have the capacity of generating conscious phenomenal experiences. But if the replacement scenario is described in this way, it is not possible to assure that the new system generates conscious phenomenal experiences in virtue of implementing the functional organization of the brain at the neural level, since the electrochemical signals transmitted by organic neurons were preserved.

We have seen, however, that artificial neurons do not need to be described as interchanging the same electrochemical signals transmitted by organic neurons. In the initial stages of the replacement process, artificial neurons need to transmit electrochemical signals in order to establish an adequate connection with organic neurons. However, when more and more artificial neurons take the place of organic ones, the use of electrochemical signals will not be necessary. However, in the case of the interchange version of the replacement thought experiment, the character of these signals was preserved in order to assure an adequate causal match with the respective new systems. For this reason, it is not clear then that there was a genuine interchange of the phenomenal and non-phenomenal states. For instance, when receiver R is attached to the physical realizers of the phenomenal states of Paul, it receives radio signals and transforms them into electrochemical signals that match the causal structure of his brain. But if we assume that these signals have a crucial role in the generation of these phenomenal experiences, Paul would not notice any change in the character of his experiences. His experiences will be exactly as before the interchange. Equally, the Chinese Nation system will not perceive any change with respect to its non-phenomenal states. There will not be any difference in the cognitive reaction of these systems to the interchange, and thus, the *reductio* strategy will be blocked.

General Conclusions

The main objective of this thesis was to present and offer a detailed evaluation of some of the most discussed versions of the replacement thought experiment in the contemporary philosophical literature. The several variations of the replacement thought experiment have been conceived as supporting a functional account of mentality, and more particularly, a functional account of phenomenal consciousness. The first versions of the replacement argument presented in the third chapter support thesis (NSS):

> (NSS) There is a set F of functional organizations such that the property of a system P of generating conscious phenomenal experiences *naturally* strongly supervenes on P instantiating a member of set F

The *regions-of-the-brain* version of the replacement thought experiment describes a device that preserves the same causal relations that a region of the brain associated with a certain mental function has with the rest of the brain. In particular, this device affects the regions responsible for the linguistic behaviour of the subject in the same way as the original organic region, and thus, his behaviour would not be modified by the replacement of the organic region. I suggested that there is indeed something problematic in the assumption that a subject that lacks conscious phenomenal experiences can behave in a way that is consistent with the presence of these experiences. However, I argued that this is not sufficient to show that the absent qualia hypothesis is false. Also, and even assuming that the phenomenal experiences of the subject were preserved by the replacement, it can be argued that the neural activity present in the replaced region, and that was associated with the generation of these experiences, was relocated to other regions of the brain. For these reasons, I concluded that the *regions-of-the-brain* version of the replacement thought experiment was not entirely satisfactory.

The neural version of the replacement thought experiment, on the other hand, provides an adequate support of thesis (NSS). We have seen that it is possible to argue that

the capacity of generating conscious phenomenal experiences is preserved along the sequence of replacement cases. There are very good empirical reasons for thinking that the generation of consciousness involves a large number of neurons working together and that it does not disappear when a single neuron in the brain of a subject is eliminated. Thus, it is extremely unlikely that consciousness disappears when a single artificial neuron takes the place of an organic one. Also, the subject cannot be aware that his conscious phenomenal experiences gradually fade along the replacement sequence. The reason is that there is no place in the brain for instantiating this awareness. Finally, if it is assumed that the conscious phenomenal experiences of the subject fade and he is not aware of it, his judgements and beliefs concerning these experiences would be systematically wrong. This clashes with the fact that the judgements that a subject makes concerning his conscious experiences are normally accurate. In particular, this assumption clashes with the Reliability Principle proposed by Chalmers, according to which our second-order judgements concerning our conscious experiences are normally correct. Since these second-order judgements can be mistaken in some cases, these principles do not have a logically necessary character. However, these principles suggest the presence of a strong empirical regularity concerning our second-order judgements. These cases exhaust all the possible ways in which the conscious phenomenal experiences of the subject disappear when the neurons in his brain are replaced by artificial ones and thus, it shows that the thesis of absent qualia is, at least, naturally false. The neural version of the replacement thought experiment also shows that there is at least one functional organization – namely, the functional organization of the brain at a neural level – such that its implementation by a physical system is sufficient for the generation of conscious phenomenal experiences. More precisely, it shows that thesis (NS) is true: there is at least one functional organization such that the property of a system *P* of generating conscious phenomenal experiences naturally supervenes on the property of *P* of instantiating this functional organization.

In the third chapter of this thesis, we have seen a further objection to the claim that the implementation of a certain functional organization by a physical system suffices for the generation of conscious phenomenal experiences. According to the thesis of Universal Instantiation, for any computer program C and any sufficiently complex physical object O,

there is a description of O under which it is implementing program C. Now, we have seen that any given functional organization can be abstracted into a CSA. This allows the formulation of the supervenience theses (NSS_M):

> (NSS_M) There is a set *C* of Combinatorial State Automata such that the property of a system *P* of generating conscious phenomenal experiences *naturally* strongly supervenes on the property of *P* of instantiating a member of set *A*.

But if the thesis of Universal Implementation is accepted, the consequence would be that any object can implement any CSA, including the CSA that corresponds to the functional organization of the brain at a neural level. Thus, thesis (NSS_M) would have the consequence that almost any object is capable of generating conscious phenomenal experiences. However, the implementation conditions of a CSA proposed by Chalmers do not allow the trivial implementations suggested by Searle. In order to implement a CSA, a physical system would need to satisfy the counterfactual transition rules of the CSA. However, we have seen that it is extremely unlikely that an arbitrary physical system could satisfy these rules. Due to these reasons, thesis of Universal Instantiation seems to be false, and thus, it does not constitute a risk for the replacement strategy.

In the fourth chapter of the thesis, I presented and evaluated a different version of the replacement argument, whose aim was to support the claim that the duplication of the functional organization of the brain by a physical system suffices not only for preserving the capacity of generating conscious phenomenal experiences, but also for the preservation of the particular qualitative character of these experiences. More precisely, this version of the replacement thought experiment supports thesis (NSS_2):

> (NSS_2) There is a set F of functional organizations such that the qualitative character of the conscious phenomenal experiences of a physical system supervenes naturally on the property of P of instantiating a member of set F.

We have seen that this version of the replacement thought experiment has been proposed for arguing against the thesis of inverted qualia, which consists in that the visual experiences generated by two isomorphic systems, P and Q, may differ in that the qualitative properties of the visual experiences generated by Q are phenomenally inverted with respect to the visual experiences generated by P. The replacement scenario of this version of the replacement thought experiment describes two beings – Zero-Paul and Ten-Paul – that differ in that there is a section *O* of the brain of Ten-Paul that has been replaced by artificial neurons, while the corresponding section *A* inside the brain of Zero-Paul remains composed by organic neurons. We assumed then that region O was implanted inside the brain of Zero-Paul along section A, in such way that the movement of a switch from position *a* to position *o* determines which of these regions is connected to the brain of Zero-Paul. It was also assumed, as a sort of *reductio* hypothesis, that the qualitative character of the visual experiences of Zero-Paul and Ten-Paul are noticeably different. Now, if the assumption were true, the visual experiences of Zero-Paul would change in front of his eyes, but he would not be able to notice this. Thus, he would not be able to form the judgement that something wrong is happening with his visual experiences when the switch is moved. This assumption clashes with the Coherence Principles suggested by Chalmers, and thus, it is extremely unlikely that the visual experiences of Zero-Paul change when the switch is moved.

Van Heuveln *et al* argue that this version of the replacement thought experiment fails. First, they claim that the movement of the switch from position *o* to position *a* does not necessarily preserve the identity of the individual. It is perfectly possible to argue that the movement of the switch generates two different beings, each of them having phenomenal experiences with different qualitative characters. Thus, no single being could have been able to experience a change concerning visual experiences, since these experiences belong to two different individuals. Second, he suggests that the generation of visual experiences involves the same mechanisms related to the recollection of these experiences, particularly those located in the visual cortex. It would be possible that the qualitative character of the visual experiences of Zero-Paul and Ten-Paul are different, but Zero-Paul would not notice

any change when the switch is moved. Thus, the *reductio* strategy of this version of the replacement argument would be blocked.

Against the objections presented by Van Heuveln *et al*, I argued first that this version of the replacement thought experiment does not need to consider a qualitative difference in the visual experiences of the subject involving more than ten per cent of the brain, and thus, it is unlikely that the replacement of this part generates two different individuals. Second, I noted that patients that suffer cortical blindness do not normally forget the qualitative character of their previous visual experiences. Thus, it is unlikely that the visual memories of an individual were so dependent on the visual cortex. Finally, I argued that assuming that a system can only experience a change in his current visual experiences by recollecting previous visual experiences is not justified. Zuboff´s version of the replacement thought experiment clearly depicts a change that would be experienced spatially and not temporally in case that the inverted qualia hypothesis was true. These reasons show that the objections presented by Van Heuveln *et al* against this version of the replacement thought experiment ultimately fails.

I consider that the reasons presented in the last section of this chapter ultimately show that interchange version of the replacement thought experiment fails. Remember that the interchange version of the replacement thought experiment was understood as providing reasons in favour of thesis (LSS):

> (LSS) There is a set F of functional organizations such that the property of a system P of generating conscious phenomenal experiences *logically* strongly supervenes on P instantiating a member of set F.

Also, if thesis (LSS) is true, then the thesis of absent qualia is logically false: a system that is a functional duplicate of a conscious being cannot lack the capacity of generating conscious phenomenal experiences. Against Van Gulick's objections to the interchange version of the replacement thought experiment, I proposed a strategy for interchanging the phenomenal and non-phenomenal states of functionally identical subjects that consisted in

attaching, to the physical realizers of these states, transmitters and receivers in order to allow an adequate causal match with the new systems. However, it is still possible that the respective phenomenal and non-phenomenal character of these states depended on the signals interchanged by their corresponding physical realizers, which are preserved because they allow a causal match with the new surrounding systems.

Remember that building artificial neurons that duplicate the signals transmitted by organic neurons seems to be perfectly conceivable. However, if the replacements are defined as transmitting exactly the same electrochemical signals transmitted by the organic neurons of Paul´s brain, we cannot be sure that the conscious phenomenal experiences of Paul are preserved in virtue of the functional organization of the system. However, the use of the same electrochemical signals is necessary only during the initial stages of the replacement process, when artificial neurons are connected to organic ones. In later stages of the process, when two artificial neurons are connected, these signals can be dispensed with.

However, that the physical realizers of the internal states of Paul and the Chinese Nation system can dispense with these electrochemical signals is not so clear. As we have seen, we need to be sure that the physical realizers of the internal states of the system causally interact with the new systems after the interchange. The only way of doing this, however, is to stipulate that the receivers installed on the realizers of these internal states transmit signals that match with the causal structure of the corresponding systems. The interchange version of the replacement thought experiment faces thus the following dilemma: if the transmission of the same input-output signals is preserved, the risk is that the physical realizers of internal states cannot match with the causal structure of the new system. But if these input-output signals are modified in such a way that these physical realizers can match with their new surrounding systems, it is difficult to see how the process produces a genuine interchange of the phenomenal and non-phenomenal states of these isomorphic systems. The reason, as we have seen, is that the signals interchanged may have a crucial role in the generation, respectively, of the phenomenal and non-phenomenal states of Paul and his Chinese Nation isomorphic system.

Bibliography

Block, N., (2007) *Consciousness, Function and Representation. Collected Papers, Vol. 1* Cambridge: MIT Press

_____ (2007b) "Troubles With Functionalism", in Block (2007) *Consciousness, Function and Representation Collected Papers. Vol. 1* Cambridge: MIT Press

_____(2002), "Searle's Arguments Against Cognitive Science", in Preston and Bishop (eds.) 2002 *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence,* New York: Oxford University Press

Block, N. *et al* (1997), *The Nature of Consciousness: Philosophical Debates*, Ed. by Ned Block, Owen J. Flanagan, Güven Güzeldere; Cambridge: MIT. Press, 1997

Ben-Menahem, (2005), *Hilary Putnam*, Cambridge University Press

Buechner, J. (2008), *Godel, Putnam, and Functionalism*, The MIT Press

Crane, T. (2003), *The Mechanical Mind*, Routledge

Cuda, T. (1985), "Against neural chauvinism" *Philosophical Studies* 48:111-27

Chalmers, D., (1995), "On Implementing a Computation", *Minds and Machines*, 4: 391-402

_____(1996), *The Conscious Mind*, Oxford University Press

_____ (1996b) "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108:309-33

_____ (2002) *Philosophy of Mind, Classical and Contemporary Readings*, Oxford University Press

_____ (2010) *The Character of Consciousness*, Oxford University Press

Dayan, P. and Abott, L.F., (2001) *Theoretical Neuroscience*, The MIT Press, London.

Egan, F. "Computation and Content", *The Philosophical Review*, Vol. 104, No. 2 (Apr., 1995), pp. 181-203

Greenberg, William M. (1998) "On Chalmer's Principle of Organizational Invariance and his 'Dancing Qualia' and 'Fading Qualia' Thought Experiments", *Journal of Consciousness Studies*, 5 (1) 1998, pp. 53-8.

Hunter, G., (1971), *Metalogic. An Introduction to the Metatheory of Standard First Order Logic* University of California Press, Berkeley and Los Angeles

Joslin, D. (2006), "Real Realization: Dennett's Real Patterns Versus Putnam's Ubiquitous Automata", *Minds and Machines* (2006) 16:29–41

Kim, J. 1984, "Concepts of Supervenience", *Philosophy and Phenomenological Research*, Vol. 45, No. 2 (Dec., 1984), pp. 153-176

_____ (1990) "Supervenience as a Philosophical Concept," reprinted in Kim (1993), 131–160.

_____ (1993) *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press

Kirk, R. (1994), *Raw Feeling*. *A Philosophical Account of the Essence of Consciousness* Oxford University Press

_____ (1996) “Strict Implication, Supervenience and Physicalism”, *Australasian Journal of Philosophy*, Vol. 74, No. 2

_____ (1999) “Why There Couldn't Be Zombies”, *Proceedings of the Aristotelian Society, Supplementary Volumes*, Vol. 73 (1999), pp. 1- 16

_____ (2005), *Zombies and Consciousness*, Clarendon Press, Oxford

_____ (2008), “The Inconceivability of Zombies”, *Philosophical Studies* (2008) 139:73–89

_____ (2001), “Nonreductive Physicalism and Strict Implication” *Australasian Journal of Philosophy*, Vol 79, No. 4, pp. 544-552,

_____ (1996) “How Physicalists Can Avoid Reductionism”, *Synthese*, Vol. 108, No. 2, pp. 157-170

Levin, Janet, "Functionalism", *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/functionalism/>.

Lewis, David. (1970) “How to define theoretical terms” *Journal of Philosophy* 67: 427–445

Majorek (2012), “Does the Brain cause conscious experience?” *Journal of Consciousness Studies,* 19, No. 3–4, 2012, pp. 121–44

McGinn, C. (1989), “Can we solve the mind-body problem?”, in Block *et al*, (1997) *The Nature of Consciousness: Philosophical Debates*, Ed. by Ned Block, Owen J. Flanagan, Güven Güzeldere; Cambridge: MIT. Press, 1997

McLaughlin, B.P., 1995. "Varieties of Supervenience," in Savellos, E. and Yalcin, U. (eds.), *Supervenience: New Essays*, Cambridge: Cambridge University Press.

Nagel, T. (1974) "What is it like to be a bat?" *Philosophical Review*, 83: 435-456.

Preston, J. and M. Bishop (eds.), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence,* New York: Oxford University Press

Pascual-Leone et al. (2005) "The Plastic Human Brain Cortex", *Annual Review of Neuroscience,* 28:377–401

Putnam, H. (1960). "Minds and Machines", reprinted in Putnam 1975, 362–385.

_____ (1967) "The Nature of Mental States", reprinted in Putnam 1975, 429–440.

_____ (1975), *Mind, Language, and Reality*, Cambridge: Cambridge University Press.

_____, Representation and Reality, The MIT Press

Rey, G., (2002), "Searle's Misunderstandings of Functionalism and Strong AI" in Preston and Bishop, (2002) *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence,* New York: Oxford University Press

Roden, C, and Karnath, H.O. (2004) "Using human brain lesions to infer function: a relic from a past era in the fMRI age?" *Nature Reviews Neuroscience* 5, 812-819 (October 2004)

Savellos, E. and Yalcin, U. (eds.), 1995. *Supervenience: New Essays*, Cambridge: Cambridge University Press.

Searle, J., (1980), ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences,* 3:417- 57

_____(1984), *Minds, Brains and Science,* Cambridge: Harvard University Press

_____ (1992), *The Rediscovery of the Mind*, MIT Press

_____(2002), ‘Twenty-one Years in the Chinese Room’, in Preston and Bishop (eds.) *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence,* New York: Oxford University Press

Shagrir, O. (1995) “The Rise and Fall of Computational Functionalism”, in Ben-Menahem, 2005, *Hilary Putnam*, Cambridge University Press, pp. 220-250

Stoljar, Daniel, "Physicalism", *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2009/entries/physicalism/.

Turing, A. M., (1936), “On computable numbers, with an application to the Entscheidungsproblem” *Proc. London Maths. Soc.,* ser. 2, 42: 230-265

_____(1950), ‘Computing Machinery and Intelligence’, *Mind,* 59: 433-460

Tye, Michael. (2006). Absent Qualia and the Mind-Body Problem, *Philosophical Review,* 115 (2): 139-168.

Van Gulick (2012). “On the Supposed Inconceivability of Absent Qualia Functional Duplicates. A Reply to Tye” *Philosophical Review,* 121 (2): 277-284.

Van Heuveln, Eric Dietrich & M. Oshima (1998). “Let's Dance! The Equivocation in Chalmers' Dancing Qualia Argument”, *Minds and Machines* 8 (2): 237-249.

Zuboff, A. (1994), “What is a Mind?”, *Midwest Studies in Philosophy*, 19 (1994).