



**Universiteit Utrecht**

DEPARTEMENT WIJSBEGEERTE

Academiejaar 2009 - 2010

JUNI 2010

**THE MECHANISM APPROACH  
APPLIED TO  
THE PHILOSOPHY OF MIND**

MATTIJS GLAS

STUDENTNUMMER: 3511367

Koekoeksplein 4A

3514TT Utrecht

t: 06 - 24 29 24 62

e: [mattijsglas@gmail.com](mailto:mattijsglas@gmail.com)

w: [mattijsglas.com](http://mattijsglas.com)

Begeleider: DR. THOMAS MÜLLER

Tweede beoordelaar: PROF. DR. MR. HERMAN PHILIPSE

Thinking about mechanisms presages new ways to handle some important philosophical concepts and problems.

- Machamer, Darden, and Craver

*Thinking about mechanisms* (2000: 23)

## TABLE OF CONTENTS

GENERAL INTRODUCTION.....	4
<b>I   HISTORICAL INTRODUCTION</b>	
1.1 EARLY MECHANISTIC THINKING AND THE ABANDONMENT OF LAWS AS EXPLANATION..	6
1.2 A SHORT NOTE ON THE USE OF INTUITION IN PHILOSOPHY.....	9
<b>II   THE MECHANISM APPROACH</b>	
2.1 THE MAIN CONCEPTS.....	11
2.2 ENTITIES AND ACTIVITIES.....	12
2.3 SET-UP CONDITIONS, TERMINATION CONDITIONS, AND INTERMEDIATE ACTIVITIES.....	13
2.4 FILLER TERMS, DIFFERENT SORTS OF MODELS AND MECHANISM SCHEMATA AND MECHANISM SKETCHES.....	14
2.5 HIERARCHIES, LEVELS AND BOTTOMING OUT.....	17
2.6 WOODWARD ON INVARIANCE UNDER INTERVENTION AND MODULARITY.....	19
2.7 ECKARDT AND POLAND ON THE CHALLENGE OF MENTAL REPRESENTATION AND PSYCHOPATHOLOGY.....	21
<b>III   THE MECHANISM APPROACH APPLIED TO FUNCTIONALISM</b>	
3.1 TROUBLES WITH FUNCTIONALISM?.....	23
3.2 FUNCTIONALISM AND THE MECHANISM APPROACH.....	25
3.3 ANIMAL AWARENESS.....	26
<b>IV   SEARLE'S CHINESE ROOM FROM A MECHANISTIC PERSPECTIVE</b>	
4.1 SEARLE'S CHINESE ROOM.....	29
4.2 MECHANISTIC LESSONS FROM THE CHINESE ROOM.....	32
4.3 CONDITIONS FOR COMPUTER CONSCIOUSNESS.....	35
CONCLUSION.....	37
REFERENCES.....	41

## GENERAL INTRODUCTION

Recently, the philosophy of science has seen the introduction of a new approach to explanation. Instead of laws, descriptions of mechanisms are taken to be explanations, especially in the diverse branches that constitute neuroscience. This so called *mechanism approach* has spawned heavy debate, and still does. As there exists a branch of philosophy called philosophy of cognitive science, and taking the relevance of philosophy for the cognitive sciences, and the other way around, into consideration,<sup>1</sup> you would expect philosophers of mind to partake in the mechanism debate, or at least to pay it its due consideration. But, talk of mechanisms is not exactly ubiquitous in the philosophy of mind. On the contrary. The term is sometimes mentioned but seldom treated with the respect paid to it by philosophers of science or scientists in the field. Thus, the leading question I will try to answer in this thesis will be if the philosophy of mind is justified in neglecting the mechanism approach. In other words, I will look for fruitful possibilities of interaction between the mechanism approach and the philosophy of mind.

To answer the leading question, this paper has broadly been given the following structure. Chapter I features a brief historical introduction. Chapter II will consist of an overview of the terminology of the mechanism approach, based on previous work done in the philosophy of science. Chapter III and chapter IV will both have the same overall structure: The first section will give an overview of an original paper, in the second section I will give an application of the mechanism approach to the original paper in question, and in the third section I will discuss the implications of the second section. Finally the conclusion will give a summary of this thesis as well as a short prospect of questions for new inquiries. To be more precise, this means the following;

We will first, in **chapter I**, take a short historical introductory tour in which we will introduce Descartes and La Mettrie as two early exponents of mechanistic thinking, next a global look will be taken at the transition from laws as explanations to mechanisms as explanations in the philosophy of science.

---

<sup>1</sup> See Bechtel (in press) for an illustration of two philosophical projects that can make theoretical and methodological contributions to cognitive science. For an outlook at a constructive attitude for philosophers willing to contribute to cognitive science, see Dennett (2009).

Then, in **chapter II**, we will take a look at the concepts involved in the mechanism approach, mainly by summarizing and explicating the landmark (2000) paper by Machamer, Darden and Craver and other relevant articles, so as to get acquainted with the terminology therein.

The following two chapters feature the original content of this thesis, and there I will cover new ground by trying to cross fertilize the mechanism approach and the philosophy of mind by the reinvestigation of two classic papers in the philosophy of mind in terms of the mechanism approach.

In **chapter III** I will take a look at Ned Block's analysis of functionalism. I will first give an outline of the discussion, after which I will show in the second section that Block's notion of liberalism can be redescribed by concepts belonging to the mechanism approach. In the third section I will give new methodological points for research into the question of animal awareness, as well as a skeptical view on that issue.

**Chapter IV** will start with an outline of J.R. Searle's attack on Strong Artificial Intelligence. After that I will, in the second section, apply the mechanism approach to this attack, and in the third section I will advance the Principle of Double Isomorphism (PDI), which I will use to show that some new answers to problems raised by Searle follow from the application of the mechanism approach and the application of the PDI. I will close off this chapter by showing that the PDI offers a new condition for the construction of machine consciousness.

The **conclusion** features a summary of the content of this thesis and sketches further possibilities for interactions between the mechanism approach and the philosophy of mind.

# I | HISTORICAL INTRODUCTION

In this chapter the modern philosophers Descartes and La Mettrie are introduced as early exponents of a mechanistic approach to the philosophy of mind. Both agree that human bodies are machines, while they differ as to the answer to the question as to whether human cognitive abilities can be explained mechanistically. Next, a look is taken at the philosophy of science, in which it used to be the case that laws were taken to be explanatory. When that assumption was questioned, a new approach evolved, an approach that will be the topic of chapter II.

## 1.1 EARLY MECHANISTIC THINKING AND THE ABANDONMENT OF LAWS AS EXPLANATION

When the French philosopher René Descartes (1596-1650) published his *Discours de la méthode* (1637), he could not have known how great an impact it would have on so many different fields. He was as clear as could be concerning the constitution of our human bodies, they are *machines*! (But, the mechanical part, the *res extensa*, would, of course, be totally helpless were it not that it has a mysterious bond with a soul, or *res cogitans*.) Descartes thus likened human bodies to machines, but not without emphasizing that the human body is '(...) incomparably better ordered and has within itself movements far more wondrous than any of those that can be invented by men.' (1998: 31 [56])<sup>2</sup>. He continues by asserting that if there existed a machine resembling '(...) a monkey or (...) some other animal that lacked reason' (1998: 31 [56]), we would not be able to determine which of the two was the machine, and which of the two was the reasonless animal.<sup>3</sup> But if there existed a machine that resembled human bodies and actions, 'we would always have two very certain means of recognizing that they were not at all, for that reason, true men.' (1998: 32 [56]) The first means he lists is to look at the use of language, for an artificial man could never use words or signs. Descartes finds it conceivable that a machine can be constructed so as to have the capacity to utter words appropriate to some circumstances or to some actions, but such a cleverly constructed machine can never engage in an actual conversation, a truly human exchange of words. The second

<sup>2</sup> Meaning: first, year of publication of the used edition, then the page-number in that publication, followed by, between square brackets, the page-number referring to the Adam and Tannery (AT) edition, which is used in most professional editions.

<sup>3</sup> The idea of a conversation as a means to determine if an entity is conscious was developed further by Turing, see: Turing (1950).

means or difference between machines and men he mentions is that such a synthetic man could perform some tasks as well or maybe even better than we, but not others. Therefore:

[I]t is for all practical purposes impossible for there to be enough different organs in a machine to make it act in all the contingencies of life in the same way as our reason makes us act. (1998: 32 [57])<sup>4</sup>

After that, Descartes claims that the rational soul can '(...) in no way be derived from the potentiality of matter, as can the other things I have spoken of, but rather that it [i.e. the soul] must be expressly created.' (1998: 33 [59]) In summary, Descartes claims that we are machines, but machines with souls which give us, amongst others, the capacity to think, to converse, and to reason: capacities or functions that machines could never perform.

A little over a hundred years later, a book, titled *L'homme machine* (1747), was published anonymously in the city of Leiden. In it, the doctor, writer, polemicist and refugee at the court of Frederic II of Prussia, Julien Offray de La Mettrie (1709-1751), takes a stand against the position of Descartes on substance dualism and his thesis that animals are soulless machines, and, instead, advocates an early version of the modern day identity theory. After insisting on comparative anatomy, behaviorism, and denying an abrupt transition between man and animal, he states that:

(...) all the soul's faculties depend so much on the specific organisation of the brain and of the whole body, that they are clearly nothing but that very organisation. (La Mettrie 1996: 26)

La Mettrie concludes his book with the words:

Let us then conclude boldly that man is a machine and that there is in the whole universe only one diversely modified substance. (...) Scholars are also the only ones whom I allow to judge the consequences which I draw from these observations, for I reject here all prejudiced men who are neither anatomists nor versed in the only philosophy that is relevant here, that of the human body. (1996: 39)

La Mettrie thus not only insists on a continuity between man and animal, but also hammers on the

---

<sup>4</sup> In all quotations, parenthesis with three dots indicate something has been left out, while everything between square brackets was not in the original and has been inserted by me.

fact that matter ('one diversely modified substance') can explain all natural phenomena including the functions allegedly performed by the soul. Although La Mettrie does argue vehemently for his monism, and was as informed about human anatomy as one could possibly be at the time, he lacked both knowledge of neural systems and knowledge of possible mechanistic realizations of mental functions, knowledge that would assist later cognitive scientists.<sup>5</sup>

Descartes and La Mettrie are two early exponents of a kind of mechanistic thinking that will be analyzed and investigated in the rest of this thesis. But discussions in the philosophy of science have, fairly recently, also picked up the notion of mechanistic thinking. Earlier though, the notion of 'explanation' in the philosophy of science was tied up with the notion of 'law'. The dominant model thereof was Hempel and Oppenheim's Deductive-Nomological (DN) model of explanation (see Hempel and Oppenheimer (1948)), which saw explanations as deductive arguments where at least one 'law of nature' had to be involved in the argument. Next to the technical internal problems the DN-model has,<sup>6</sup> it also got opposed for different reasons, one of which was the claim that laws are not explanatory.<sup>7</sup> As Cummins stated it:

No laws are explanatory in the sense required by DN. Laws simply tell us what happens; they do not tell us why or how. (...) Surely the correct moral to draw here is that [laws are] an *explanandum*, not an *explanans*. (Cummins 2000: 119)<sup>8</sup>

Thus, since laws were no longer seen as explanatory, this opened up a new philosophical field, referred to here as the mechanism approach. Peter Railton (1978) (cf. Glennan 2002: 342) was the first to introduce the idea of mechanisms into the debate on explanation. Railton, while keeping the nomological aspect of explanation, also stated that an account of the workings of the mechanism involved, needed to be added to the explanation. But Railton remained vague on what exactly he took a mechanism to be. Only much later intricate accounts of what mechanisms could be taken to be appeared, most notably, 'Thinking about mechanisms' (2000) by the philosophers Peter Machamer, Lindley Darden, and Carl F. Craver. The next chapter will feature a thorough explication of that paper, so as to familiarize the reader with the different concepts concerning the mechanism approach in the philosophy of science, while at the same time getting clear about their meanings

---

5 For a longer discussion of Descartes and La Mettrie's place in the larger historical picture, see Wright and Bechtel (in press). See also Machamer, Darden and Craver (2000), pp. 14-15.

6 E.g. the extent to which the DN model can be applied to, for instance, neuroscience; what laws are and whether they exist in nature.

7 At least, not in scientific fields such as psychology.

8 In this quotation, as in all others, emphasis are as they are in the original, unless explicitly stated otherwise.

and interactions.

But first, in the next section, we will briefly discuss the use of intuition in philosophy, as some philosophers (e.g. Jim Woodward, as we will see in section 2.6, and Ned Block, as we will see in chapter III) use intuitions, amongst other purposes, to assert their philosophical theses.

## 1.2 A SHORT NOTE ON THE USE OF INTUITION IN PHILOSOPHY

Of course, the general public has an affinity for notions such as 'soul' and 'mind', and is antagonistic to scientific, let alone mechanistic accounts of mental abilities. Very often this is because the scientific and philosophical accounts are highly counterintuitive.<sup>9</sup> But, as we will see below (in section 2.6 and chapter III), philosophers also often refer to intuitions, and use them to support their arguments. But this can be a very dangerous habit in scientific as well as philosophical matters. Hopefully, the following short, typically Wittgensteinian, anecdote will serve to illustrate a minor point I wish to make on the topic of intuition and the guiding role it has been assigned in philosophy's quest for good theories.

He [Ludwig Wittgenstein] once greeted me with the question: "Why do people say that it was natural to think that the sun went round the earth rather than that the earth turned on its axis?" I replied: "I suppose, because it looked as if the sun went round the earth." "Well," he asked, "what would it have looked like if it had *looked* as if the earth turned on its axis?" (Anscombe, E., quoted in Metzinger 2009: vi)

Often, theories are criticized because of their counterintuitiveness. For example, a well-built theory of intentionality might, perhaps, have the consequence that some sort of machine is included in its category of intentional creatures. And because of the counterintuitiveness of that consequence, the theory is then criticized or simply cast aside as nonsense. A second example would be the Gettier-objection to the knowledge as a justified-true-belief (JTB) theory. Gettier gave a short (and I think, ultimately misguided) example of the consequences of the JTB-theory. His argument was basically a *reductio ad absurdum*. Because of the counterintuitive consequence he showed, he regarded the JTB-theory to be wrong. But an important question a philosopher should ask himself in such cases is if he is not better off by simply biting the bullet. Of course some philosophical or even scientific theories (e.g. quantum mechanics) are bound to have counterintuitive results. But can such results be held against a theory? I should emphatically say that they cannot. They can, instead, be seen as an argument against the presumed victoriousness or correctness of our intuitions. Just like ancient

<sup>9</sup> With the term 'intuitions' I am here referring to common sense ideas about how the world works.

people who presumed the sun to be revolving around the earth instead of the other way around, were simply having misguided, uncriticized intuitions, a lot of modern intuitions are likely to be challenged as the result of new theories in different philosophical and / or scientific fields. If one wants to save his intuitions from those challenges, one should simply stop reflecting on the numerous questions asked about numerous other questions and the answers given to answer those questions. As D.C. Dennett stated: 'Any theory that makes progress is bound to be *initially* counterintuitive.' (1989: 6) The quest for theories that conform to our intuitions should be a contrafactic quest, a guiding search heuristic, but not a dogma, not a criterion to which scientific / philosophical progress should be subjected. Where intuitions keep on failing, we should ultimately, but not immediately, abandon them, and let progress commence.

## II | THE MECHANISM APPROACH

In this chapter, which can be seen as the groundwork for the rest of this thesis, the different concepts concerning the mechanism approach in the philosophy of science will be described and elucidated, so as to get clear about their meanings, interactions, and possible uses in the next chapters.

Machamer, Darden and Craven (2000)<sup>10</sup> can already be called the *locus classicus* concerning scientifically/philosophically informed talk about mechanisms. They propose that '(...) much of the practice of science can be understood in terms of the discovery and description of mechanisms.' (MDC, 2000:2) And because they hold the position that as of up to the time that they wrote their paper, an adequate analysis of what mechanisms are and how they work in science was lacking, they propose to sketch a mechanistic approach to two scientific fields; neurobiology and molecular biology. In the chapters III and IV we will be concerned with the question what the results of an application of the mechanistic approach to the philosophy of mind would be. But first we have to take a look at what philosophers of science have said about the concepts of the mechanism approach.

### 2.1 THE MAIN CONCEPTS

MDC divide a mechanism into two basic parts: *entities* and *activities*, which, together, play a certain role called a *function*. The interactions of the entities concerned proceed in a *regular* fashion, and have *productive continuity*, which *bottoms out*, meaning that the entities are composed of other, lower-level mechanisms, the latter of which the scientists concerned with the upper level-mechanisms are not bothered with. The function performed by the mechanism lies between the boundaries of the *set-up* and the *termination conditions*, in between of which lie *intermediate activities*.

On a higher level, mechanisms are nested, which means they occur in *hierarchies* of higher- and lower level mechanisms. Proposed mechanisms can be subdivided into *mechanism schemata* and *mechanism sketches*, the intelligibility of which can be emphasized by the use of three terms:

<sup>10</sup> From now on, for reasons of brevity, references to Machamer, Darden, and Craver (2000) will be given as MDC.

from *how-possibly*, via *how-plausibly*, to *how-actually models*. In the following paragraphs the meaning of the terms mentioned will be elucidated, in order to (1) lay down the groundwork for the following chapters and (2) provide a basic introduction to the mechanism approach for readers unfamiliar with the subject.

## 2.2 ENTITIES AND ACTIVITIES

MDC provide the following abbreviated definition of what a mechanism is:

Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (2000: 3)<sup>11</sup>

They also claim that someone who provides a description of a mechanism for a phenomenon *x*, is thereby *explaining* the phenomenon *x*.

An example of a mechanism explaining a phenomenon, though outdated, is Hebb's rule, which purports to explain some types of associative learning (or, Hebbian learning, where something is learned through its association with something else, as in the standard example of the goldfish that comes up to the surface of the water to find food, because it has associated the tingling of a bell with the appearance of food) in virtue of the increased effectiveness of axons stimulating connected cells, where activities (in this case the increased efficiency) are the producers of change, and entities (in this case the axon and the connected cell) are the things that engage in the activities. The function performed in this example would be associative learning.<sup>12</sup>

As stated above, a mechanism should also exhibit regularity, it should work in the same way under the same conditions.<sup>13</sup> The productive continuity of a mechanism consists in the presence of ways in which the steps between the different stages are connected. Stages and intermediate activities can be represented schematically as 'A→B', where the letters denote the stages and the arrows denote the gaps that need to be bridged by an activity (MDC 2000: 3). If a precise description of the activity bridging the gap between a stage A and B is lacking, the description is not complete. In the case of Hebb's rule as a mechanism for explaining the phenomenon of associative learning, the schema for the start or set-up conditions would be 'A→B', and the schema for the

11 Cf. Glennan (2002: 344) 'A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.'

12 Hebb's rule is here strictly used as an example. In their (2000), MDC use a different example, namely Long-Term Potentiation (LTP), to illustrate their mechanism approach. The choice to use a different example was made for the didactic purpose of acquainting myself thoroughly with these concepts, where the underlying assumption is that the application of a concept to new examples enables one to really understand them.

13 This is not generally accepted. See Woodward (2002) and the discussion of that paper in section 2.6 below.

finish or termination conditions could be either (1) ' $A_c \rightarrow B$ ', (2) ' $A \rightarrow B_c$ ', (3) ' $A \rightarrow cB$ ' or (4) ' $A_c \rightarrow B_c$ ', where the subscript  $c$  denotes change. Hebb's original formulation ran:

*When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased. (Hebb 1949: 62)*

This formulation has the four consequences schematically described above. Which one of the four turned out to be correct (if any) need not bother us here since we are using Hebb's rule strictly for illustratory purposes. The important thing is that Hebb mentions both set-up and termination conditions that can be confirmed experimentally.

### **2.3 SET-UP CONDITIONS, TERMINATION CONDITIONS, AND INTERMEDIATE ACTIVITIES**

The description of a mechanism tends to start with an idealized description of the set-up conditions, meaning the relevant entities and their properties. And so does the description of the termination conditions (MDC: 11-13). Most of the features in the set-up conditions do not contribute inputs to the mechanism, but are entities within the mechanism. The termination conditions are 'a privileged endpoint' as they describe an equilibrium, an activation or inhibition, or the production of a product. MDC prefers to describe what are usually called 'outputs' as 'termination conditions', because that term emphasizes the fact that in most descriptions of mechanisms, there is nothing that 'comes out' of the mechanisms. The description of the entire mechanism is an idealization, so the word 'conditions' is meant to underline the referential aspect of the description. As stated before, mechanisms are most often nested into a hierarchy of other higher- and lower-order mechanisms, and it is not unusual at all for a mechanism to constitute a single stage in another, larger mechanism.

The intermediate activities are to be seen as descriptions of the activities that take the mechanism from the set-up to the termination conditions. Such a description has to show how '(...) the actions at one stage affect and effect those at successive stages.' (MDC: 12) In the case of Hebbian learning, the intermediate activities would be the activities following the firing of a cell, which should finally end up changing 'the effectivity of the firing cell', one way or another.

## 2.4 FILLER TERMS, DIFFERENT SORTS OF MODELS AND MECHANISM SCHEMATA AND MECHANISM SKETCHES

As the reader may notice, the last sentence of the previous section contains quite a lot of obscure words. Obscure because they lack any further definition. Examples would be the words 'firing', 'cell', 'changing' and 'effectivity'. Craver (2006: 360) calls such obscure words *filler terms*, because they fill a void with a word instead of a (complete) description, without explicitly admitting such an act. Filler terms thus fill the 'gaps' in the schematic representations of mechanisms described above, and '(...) are barriers to progress when they veil failures of understanding.' (Craver 2006: 360). Thus, close scrutiny on behalf of both philosophers and scientists is necessary to avoid the 'illusion of understanding' associated with filler terms.

When it is the case that no filler terms are used in the description of a mechanism, the further question is how much understanding the description conveys. As scientists are usually highly specialized, not everybody can be aware of everything happening in other fields. A coevolutionary strategy (Churchland and Sejnowski 1988: 741), whereby researchers in different domains share and depend on each others knowledge and theories, is thus necessary for understanding. But, this also means a certain use of filler terms cannot be avoided. For, because of the dependence on each others knowledge and theories, scientists are bound to use terms they themselves do not know everything, or maybe just a little, about.

When all appropriate conditions for the description are met, Craver would speak of an 'ideally complete description of a mechanism'. Such a description would include 'all the entities, properties, activities, and organizational features that are relevant to every aspect of the phenomenon to be explained.' (Craver 2006: 360) This is supposed to be a contrafactic<sup>14</sup> description. Everybody wants it, but nobody actually makes one: it would contain too many unnecessary and / or irrelevant details.

In between on the one hand, a *mechanism sketch*, meaning a partially incomplete description of a mechanism, and, on the other hand, an ideally complete description, lie what Craver dubs '*mechanism schemata*'. Those are descriptions that are more or less abstracted from, in Philip Kitchers words, the 'gory details'. The meaning of the difference between a mechanism sketch and a mechanism schema is conveyed in the meaning of the two words 'sketch' and 'schema'. Sketch should imply what it usually implies, that one has an idea of the final answer, but not a definite conclusion. On the other hand, a schema is already based on some research or thinking, but still has

---

<sup>14</sup> Not to be confused with 'contrafactual', the word 'contrafactic' denotes a struggle to attain something, even though it is generally accepted that it is unattainable, thus, an ideally complete description is 'contrafactic' in the sense that it is practically unattainable, even though it is stated to be the ultimate goal.

gaps.

Craver (2006) also describes the ways in which descriptions can vary in their '*mechanistic plausibility*.' A description of a mechanism for a phenomenon *P* should not just describe a way to get from the set-up to the termination conditions, but should be a description of the way the mechanism *in the phenomenon P* gets from beginning to end. So, in the search for explanations, a mere simulation of the input-output behavior of a phenomenon is not enough. The plausibility of the mechanism descriptions on the continuum ranging from mechanism sketches via mechanism schemata to ideally complete descriptions, can, it is proposed by Craver, be captured by a division of on the one hand How-Possibly (Hpo) models, on the other hand, How-Actually (Ha) models, and How-Plausibly (Hpl) models in between.<sup>15</sup>

Hpo models are not to be seen as explanations, but rather as clues towards the right direction in search of the correct explanation. Craver thus states Hpo models are '(..) often heuristically useful in constructing a space of possible mechanisms,' (Craver 2006: 361). In the case of Hebbian learning, the four ways in which Hebb's hypothesis can be correct, can be seen as Hpo models, although they do not rise above the level of flow-chart or black-box series. But the important point is, as Craver stated, that they narrow down the search space.

Hpl models are, by definition, empirically better informed than Hpo models. They are constrained by the knowledge available in the field. A Hpl model in the case of Hebbian learning would thus be one of the four schemata mentioned, although a Hpl model would lack further details. If scientists found out the exact way in which Hebbian learning occurs, if they could fill in all the gaps in the mechanism schemata, a Ha model would be the result. Ha models describe the actual details of the mechanism under scrutiny, and so the entities, activities, intermediate activities, set-up and termination conditions in a phenomenon *P* should have isomorphically identical counterparts in the Ha model.

Craver has also proposed certain criteria for assessing mechanistic explanations in a normative manner. These criteria are, at the same time, helpful to distinguish Ha from Hpl models (Craver 2006: 268-369). Those criteria are the following;

As phenomena are typically multifaceted, a good description should include (1) the *different features* associated with the phenomenon in question. A good description should also address the (2) *precipitating conditions*, as it is often the case that the termination conditions can be produced through different mechanical pathways. The (3) *inhibiting conditions*, the circumstances under

---

<sup>15</sup> See also: MDC: 21, who state: 'Mechanism descriptions show *how possibly*, *how plausibly*, or *how actually* things work. Intelligibility arises not from an explanation's correctness, but rather from an elucidative relation between the explanans (the set-up conditions and intermediate entities and activities) and the explanandum (the termination conditions or the phenomenon to be explained).'

which the phenomenon does not occur, should be emphasized because when one truly understands a mechanism, one should be able to say why it does not work under certain conditions. To know the (4) *modulating conditions* is to know how variations in background conditions change the way in which the mechanism works.

Laboratory conditions are just that: laboratory conditions, and thus non-standard. Although they might seem unnatural in that respect, it is important not to neglect the ways in which mechanisms act when manipulated, even if the manipulation is of an artificial nature. So (5) *non-standard conditions* should also be taken into account. Also, the study of (6) *by-products* can deliver important information to distinguish HPO from Ha models, because although byproducts have no function in the phenomenon under study, they do deliver concrete information about the larger scale relation of the phenomenon to other entities and / or processes. If more than one mechanism can account for the phenomenon, the by-products can be used to distinguish the Hpo from the Ha-model.

Further, MDC mention a few other theoretically important functions of a mechanism schemata. According to them, mechanism schemata play the role often attributed to theories:

They are used to describe, predict, and explain phenomena, to design experiments, and to interpret experimental results. Thinking about mechanisms as composed of entities and activities provides resources for thinking about strategies for scientific change. (MDC: 17)

Accordingly, scientists can use mechanism schemata to build complex mechanism sketches (cf. Dennett 1981: 123-4). MDC also state that mechanism schemata can be used to yield predictions, and to provide blueprints for designing research protocols. In the case of predictions, these can be either borne out, or these can fail to do so. If a prediction is borne out, this constitutes a proof for the appropriateness of the schema, and if the prediction fails to come true, this provides anomalies which can then be used as the starting point for new research. For example, assumptions about the set-up or termination conditions could be wrong, so a scientist can alter them to see what happens then. Using the results of those experiments, the mechanism schema can be either altered, or ultimately, abandoned altogether.

In the case of blueprints, a scientist can design an experiment in which a physical instantiation of the schema is chosen, upon which interventions can be conducted, so that the scientist can find out more about the behavior of the mechanism under different circumstances. This can also constitute evidence for the schema under investigation. If, for example, certain changes that the scientist notices in the experiment can also be seen in the real life behavior of the

phenomenon which the schema is about, this would be a clue to the correctness of the schema (MDC: 16-18).

## 2.5 HIERARCHIES, LEVELS AND BOTTOMING OUT

MDC describe mechanisms as being nested in hierarchies, with which they mean that (e.g.) a box in one mechanism schema can be substituted by a description of another (complex) mechanism. Depending on the level of analysis, those boxes are or are not objects of research. This can be helpfully illustrated by a division made by Churchland and Sejnowski (1988). They distinguish three different notions of the terms 'level' (741-743). First, there are levels as levels of *analysis*, by which they mean the 'conceptual division of a phenomenon in terms of different classes of questions that can be asked about it' (1988: 741); levels as levels of *organization*, which concerns the question how many different scales of organization (e.g. molecules, synapses, neurons, networks) are involved, and, thirdly, as levels of *processing*, which involves looking at the different tasks performed within a given function. The first notion of level mentioned above concerns us here, for it clarifies that many different research questions can be asked about a single phenomenon.<sup>16</sup> Thus, some things that some scientists take for granted are under research by other scientists, like chemists who take the existence and functioning of atoms for granted, while physicists are still investigating the constitution of atoms. Another example is the way in which evolutionary psychologists ask themselves questions regarding the evolutionary rationale behind certain mental functions or phenomena, while taking the mental functions themselves (e.g. greed, shame, homosexuality, linguistic skills) for granted. Evolutionary psychologists are not researching the physical basis of the phenomena, or asking themselves which parts of the brain are necessary or sufficient for the exercise of those functions (which is a task of neurophysiologists), or which neurons have to do what for a human brain to be able to feel shame (which is a task of neurologists). The three scientific disciplines mentioned last each take the discipline 'above' them and the one 'below' them for granted. MDC state that the point where research bottoms out is the point where '(...) the components are accepted as relatively fundamental or taken to be unproblematic for the purpose of a given scientist, research group or field.' (MDC: 13) The meaning of the concept of bottoming out is thus slightly misleadingly called '*bottoming out*', because a given phenomenon or mechanism can also 'bottom out' at the top or, (thus) 'top out'.

---

16 But, as Craver (2001: 63) states: 'Exactly how many levels there are and how they are to be individuated are empirical questions that are often answered differently for different phenomena.' (Although, as a philosopher, I would not say that the question how many levels there are can be answered solely by empirical inquiry, because, as Churchland and Sejnowski's (1988) analysis makes clear, it is also a theoretical / philosophical question / problem.) See also: Bechtel and Herschbach (in press (Philosophy of the cognitive sciences): 10-12).

Craver (2001) thinks that the mechanism approach sheds a light on the question of *interlevel integration*, where interlevel integration can be said to have the goal of uniting descriptions of different levels into one coherent mechanism. (Craver 2001: 63) According to Craver, the mechanistic hierarchy that the mechanism approach entails, solves the problem of levels. The problem of levels being the problem of how to look at the different levels of nature. Levels are not, in Craver's words, 'divisions in the furniture of the world', (2001: 67) one who commits himself to different levels does not have to reify levels, no division of nature is necessary.

Craver also asserts that there is a difference between *contextual* and *isolated* descriptions of the system or its function. For example, an isolated description of a nerve would describe it as a pathway for electrochemical impulses to travel through, where a contextual description would pay attention to the sort of signals (e.g. signals for proprioception) the nerve delivers, and what role the message that the signal is, plays in the adjacent system(s). Thus, an isolated description '(...) draws an idealized dividing line at the spatial boundary of the item and recognizes a limited number of crucial interfaces across that otherwise closed boundary.' (Craver 2001: 64) Where a contextual description describes the system and its capacity '(...) in terms of its contribution to a higher (...) level mechanism. The description includes references not just to X (and its  $\phi$ -ing) but also to X's place in the organization of S's  $\psi$ -ing.'<sup>17</sup> (Craver 2001: 63)

Summarizing the last paragraph, there is a three-way division into alternative but complementary ways to describe an S's  $\psi$ -ing. First off, in *isolation* (0), secondly, *constitutively* (-1) when looking down to lower-level mechanisms, and last, *contextually* (+1) when looking up to higher-level mechanisms. The important point is that these three ways of describing a system's capacity are '(...) three different *perspectives* on that item's activity in a hierarchically organized mechanism; they are not levels of *nature*.' (Craver 2001: 66) This perspectivalist approach to levels thus sees the world as a mechanistic hierarchy, which can be divided, but the division is relative to '(...) a perspective on an activity at a given level in a mechanistic hierarchy.' Interlevel integration will then be achieved when one knows how these three perspectives complement each other, or, in other words, when one knows what the  $\psi$  of S is, how S works, and how the X's  $\phi$ .

---

Now that the terminology of the mechanism approach in the philosophy of science has been discussed without any reference to current debates, it is time to discuss two challenges concerning

<sup>17</sup> Where: S = system;  $\psi$  = the capacity of S; X = a component of S;  $\phi$  = the capacity of X. In the example of the nerve and its function, S would be the brain and X would be the nerve, where  $\psi$  would (in the case of proprioception) be the brain's knowledge of the position of its limbs, and  $\phi$  the capacity of the nerve to transmit signals. And of course, (see also the discussion of Churchland and Sejnowski (1988) above) different research questions can be asked about an S's  $\psi$ .

the mechanism approach, so as to emphasize that not everybody agrees on everything concerning the mechanism approach. Hopefully, this will give the reader some insight into the shortcomings of the mechanism approach. The first challenge is that of counterfactuals, raised by Jim Woodward (2002), the second challenge is that of mental representation and of psychopathology, raised by Barbara von Eckardt and Jeffrey S. Poland (2004).

## 2.6 WOODWARD ON INVARIANCE UNDER INTERVENTION AND MODULARITY

In his paper *Mechanisms: A counterfactual account* (2002) Woodward proposes a counterfactual account of what mechanisms are. His proposal hinges on the notions of (1) *invariance under intervention* and (2) *modularity*, each of which, or so he proposes, can be captured by certain counterfactuals. With his proposal, Woodward is trying to give '(...) a more general and less discipline specific characterization of notions like "mechanisms" and "production."' (2002: 367)

The notion of invariance under intervention is used by Woodward to clarify what the regular productive behavior of a mechanism or a component thereof is. The intuitive idea he tries to capture is that productive behavior is like lawlike behavior. Only, the notion of law is not applicable to the workings of mechanisms, because the behavior of most mechanisms is far from law-like:

There is a major mismatch between the features that philosophers have thought laws must possess and the generalizations that characterize the operations of many mechanisms. (2002: 368)<sup>18</sup>

For example, mechanisms usually do not work in circumstances other than 'normal' (e.g. a gun cannot fire when it is submerged in water), so their behavior cannot be called 'general', which would be a necessary (but not sufficient) condition for calling the mechanisms' behavior lawlike. So a description is needed that distinguishes lawlike from mechanical behavior. Thus, with invariance under intervention Woodward tries to substitute a productive relationship for a causal relationship, where 'productive' means that a mechanism is a mechanism for something. If it did not 'do', anything, it would not be a mechanism, as is obvious from the definition of mechanisms by MDC, which speaks of entities and *activities*. This activity of the mechanism, the thing it does, is what Woodward is trying to elucidate with his notion of invariance under intervention. The notion of invariance under intervention is thus meant to capture the intuitive idea that when one is trying to figure out if X causes Y, the mechanism (X) has to maintain its productive behavior (X's production of Y) under some but (not all) interventions, if it is to be possible to regard the mechanism as the

<sup>18</sup> See also the section 1.1 above

cause of the thing produced (2002: 369-371). If that is the case, if the relationship is indeed invariant under certain conditions, it should be possible to manipulate and / or control the relationship (i.e. the productive behavior).

This proposal of Woodward is motivated by his conviction (along with, for example molecular biologists like Robert Weinberg (1985: 48)) that ever since molecular biology is able not just to describe nature but to actually intervene in its behavior, the field has become explanatory instead of just descriptive.<sup>19</sup>

Woodward's second idea is the condition of modularity, which he defines as the idea that:

(...) the components of a mechanism should be independent in the sense that it should be possible in principle to intervene[,] to change or interfere with the behavior of one component without necessarily interfering with the behavior of others. (Woodward 2002: 374)

For example, take a system  $S$  which is composed of components  $X_1...X_n$  and a process  $P$  that is capable of altering the behavior of  $X_1$ . If  $S$  is modular and it is known what the behavior of  $X_1$  under the influence of  $P$  will be, then this knowledge, combined with the knowledge about the unchanged behavior of the other components  $X_n$ , can be used to predict the new behavior of  $S$  as a whole. E.g., when sugar is put in the oil tank of a car, we can predict that the car will stop moving because the engine will fail although the rest of the parts of the car will remain unaffected. Or, to take a biological example, when someone's corpus callosum is severed, we know that he will still be able to see, to walk, to talk and so on, while he will no longer be able to vocally name an object seen within the left visual field of his brain (when, as is usual, the speech-control center is in the left side of the brain), because the visual information is only sent to the right side of the brain. The condition of modularity is meant, by Woodward, to be used to determine if a decomposition of a mechanism into parts is correct. If the modularity condition holds, the decomposition is indeed correct, and if the condition does not hold, the decomposition was incorrect. So, to take another example, if the activities of the primary visual cortex (V1) of the human brain, an area that is said to act as the pattern recognition device of the perceptual system, was disrupted by electronic impulses, and the subject under investigation would suddenly, next to his ability to discriminate different patterns,

---

<sup>19</sup> See also the (2006) paper by Craver discussed above, which also holds (among others thing) that mechanistic models are explanatory because of the ability to intervene and / or control: '(...) explanations outperform models that are merely phenomenally adequate because they (...) afford a greater possibility of control over the phenomenon, and so allows one to answer a greater range of questions (...)' (Craver 2006: 358)

also lose his ability to speak English, it should be said that the identification of V1 as a 'pattern recognition device' was wrong, which should lead to new investigations to figure out what role V1 plays in the brain's speech-producing system.

Finally, Woodward proposes, amongst other things, the following criteria to determine if a representation can be seen as an acceptable model of a mechanism. First, the representation needs to describe an organized set of components, where, secondly, the behavior of each component in the description should be invariant under certain conditions, and, finally, the generalizations governing the different components should be independently changeable. The former should, according to Woodward, be seen as a simplification of the general idea that '(...) components should be independently changeable.' (Woodward 2002: 375)

## **2.7 ECKARDT AND POLAND ON THE CHALLENGE OF MENTAL REPRESENTATION AND PSYCHOPATHOLOGY**

Barbara von Eckardt and Jeffrey S. Poland's (2004) investigation of MDC argues that the mechanism approach '(...) cannot completely capture all aspects of the content and significance of mental representation or the evaluative features constitutive of psychopathology.' They claim this because of the externality of certain factors important to a complete description or treatment of the phenomena mentioned. In the case of the content and significance of mental representation, they claim that the mechanism approach fails to capture the externalistic aspects involved, because many theories that try to explain how representations are naturalistically realized are of an externalistic nature, meaning that '(...) the naturalistic properties and relations of a representation bearer that realize its content are taken to be properties and relations that extend beyond the head (...)' (2004: 981) According to Eckardt and Poland it is clear that externalistic components cannot be part of a neural mechanism.

Likewise, in the case of psychopathology, they argue that because psychopathological phenomena are normatively classified, a constitutive explanation of those phenomena '(...) will involve a comparison of the cognition or behavior with a norm and determine whether the norm is statistical, based on natural design, based on community standards, or epistemic'. (2004: 982) Such a comparison will have to involve more than just a description of the mechanism.

Summarizing the above, Eckardt and Poland claim that constitutive explanations in the neurosciences have to involve more than just descriptions of mechanisms. In the case of mental representations a correct explanation needs to address the content and significance of the representations, which will not lie within the boundaries of the described mechanism. And in the

case of psychopathology they argue that accounts of types of dysfunctions will need an explanation as to why the cognition or behavior may be given the evaluative label it has, and such an explanation requires a comparison with norms which lie outside of the boundaries of mechanism hierarchies.

### III | FUNCTIONALISM, THE MECHANISM APPROACH, AND ANIMAL AWARENESS

Now that we have covered the groundwork for this thesis by explicating the terminology of the mechanism approach in the previous chapter, we will move on to the question of the relevance of the mechanism approach to the philosophy of mind. We will do this by the reexamination of Ned Block's famous article 'Troubles with functionalism' (1980).<sup>20</sup> I will try to shed some new light on the issues raised in that text, as we have with us a new set of ideas; the ideas and concepts of the mechanism approach. The first section will give an outline of the argumentation of the article itself. In the second section, I will cover new ground by applying the mechanism approach to this classic text, which will lead to some new thoughts on the investigation of animal awareness in section 3.

#### 3.1 TROUBLES WITH FUNCTIONALISM?

In his (1980) Block delivers a critical analysis of functionalism, which is defined by Block as the thesis that:

(...) each type of mental state is a state consisting of a disposition to act in certain ways *and to have certain mental states*, given certain sensory inputs and certain mental states. (1980: 268)

He then distinguishes two forms of functionalism: Functionalism and Psychofunctionalism,<sup>21</sup> where Functionalism treats functional analyses as analyses of the meaning of mental terms, and Psychofunctionalism treats functional analyses as 'substantive scientific hypotheses' (1980: 271-272). This twofold separation is meant to emphasize that the difference between the two forms is a difference between input-output specifications. Functionalists can only use input-output specifications that can be classified as 'part of common-sense knowledge', where Psychofunctio-

<sup>20</sup> Originally published in 1978, but all references are to the reprint in Block (1980).

<sup>21</sup> In the sections concerning Block's (1980), we will follow him in writing Functionalism and Psychofunctionalism with capitals when distinguishing between the two currents just defined, and using a lowercase for talk of functionalism that is neutral between the two currents.

nalists have the freedom to use internal parameters, e.g. signals transferred between neurons. Furthermore, Block levels two accusations against functionalism. The first accusation, that of *liberalism*, holds that functionalism classifies '(...) systems that lack mentality as having mentality' (Block 1980: 275 (see also: 269), while the second accusation, that of *chauvinism*, holds that some forms of functionalism are so stringent they withhold mental properties from systems that in fact have mental properties (1980: 270).

Block argues that functionalism is guilty of liberalism, that functionalism can be protected from the chauvinism critique by modifying it to embrace empirical psychology (thus producing Psychofunctionalism), and that no version of functionalism is able to resist the liberalism as well as the chauvinism critique.

To illustrate the accusation of liberalism, Block uses two thought-experiments. The first describes a body like human bodies, the head of which is internally regulated by homunculi who perform the tasks necessary to simulate the input-output functions that describe humans. According to Functionalism, this homunculi-system, which is, *ex hypothesi*, functionally equivalent to humans, does have mentality just like regular humans. But, as Block states, most readers will have the strong intuition that the homunculi-system lacks mentality, that it does not feel, does not have a mental life like normal humans do, and he then argues that that intuition has a rational basis, which constitutes a good reason for the claim that Functionalism (but not necessarily Psychofunctionalism too) is false.<sup>22</sup> Block then tries to show why different attempts to rescue Functionalism from this first thought experiment will fail.

One attempt to save Functionalism from the homunculi-headed counterexample adds an extra safety measure to Functionalism, by stipulating that '(...) two systems cannot be functionally equivalent if one contains parts with functional organizations characteristic of sentient beings and the other does not.' (1980: 279) But, according to Block, there is also a counterexample to that rescue attempt. For consider intelligent aliens who are (in comparison to us) extremely small, not just microscopically small, but microphysically small. They discover our part of the universe and the physical constitution thereof, and start to rebuild, from spaceships and alien-matter, elementary particles which are functionally identical to the particles in our universe. We, humans, then get mixed up with the functionally identical particles and eventually end up consisting completely of alien matter (1980: 280). Now, what is the difference between the elementary-particle-people and the homunculi-headed example? According to Block, the difference, functionally seen, is immense:

---

<sup>22</sup> Cf. section 1.2 above.

As far as 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 matter. (...) What seems important is *how* the mentality of the parts contributes to the functioning of the whole. (1980: 280)

In both examples, homunculi invade our system, but in the homunculi-headed example, the homunculi are of a noticeable bigger size than in the elementary-particle-people example. A simple MRI scan could show the difference. As we become homunculi-infested in the elementary-particle-people way, this does not make a difference to "(...) your psychological processing (i.e. information processing) or neurological processing but only to your microphysics." (1980: 280)

Thus far, Block has attempted to show how Functionalism fails. Its boundaries are too wide, for even something that mimics human input-output functions will be described by Functionalism as having mentality. But what if, as a rescue matter, we include neuropsychological theories in the Functionalism theory, so to arrive at what Block dubbed Psychofunctionalism. This could be done by adding the measure that:

(...) it is logically impossible for a system to have beliefs, desires, etc. except insofar as psychological theories true of us are true of it. Psychofunctionalism (so understood) stipulates that Psychofunctional equivalence to us is necessary for mentality. (1980: 291)

But this stipulation, according to Block, succumbs to chauvinism. For if we were to meet aliens who are Functionally equivalent to us, while Psychofunctionally very different, they would, according to Psychofunctionalism, have no mentality, simply because they lack Psychofunctionalist equivalence. So counterexamples can be given both to Functionalism and Psychofunctionalism, which shows, at least according to Block, that both need to be rejected.

### **3.2 FUNCTIONALISM AND THE MECHANISM APPROACH**

It is interesting to see that, in his analysis, Block developed a few notions that can readily be analyzed by the mechanism approach. For short, where Block remarks that our psychological processing would not be affected by the change in microphysics in the elementary-particles-people example, this readily summons the notion of bottoming out, where scientists accept the components they deal with as '(...) unproblematic for the purpose of a given scientist, research group or field.' (MDC 2000: 13)

Further, Block's notion of liberalism, the denial of mentality of systems that have mentality,

as applied to theories about how minds can be analyzed, can be redescribed in mechanistic vocabulary as a struggle between How Possibly (Hpo), How Plausibly (Hpl) and How Actually (Ha) models of the working of the mind. An application of the mechanism approach to this issue can hopefully, as Craver states, provide new criteria:

Perhaps grounding functional description in the details of mechanistic organization will provide a set of criteria for assessing the precision and accuracy of functional ascriptions and will perhaps help to guard against empirically inadequate, vague, or overly abstract functional ascriptions. (2006: 73)

It has to be noticed that where Craver speaks of 'precision and accuracy', this too is relative to the view of the scientist and thus relative to different scientific points of view. So, where Block states that Functionalism is too liberal because it ascribes mentality to a homunculi-headed system of which our intuitions state it does not have mentality, and where he states that Psychofunctionalism is too chauvinistic because it would deny mentality to aliens who obviously have it, in terms of the mechanism approach, the following analysis can be applied.

If it were said of Psychofunctionalism that it provided a Ha-model of humans as well as of cognition in general, this is false, for few other creatures (maybe not even monkeys) have exactly the same mechanistic structure as the human brain has. If it is said of Functionalism that it provides a Hpl model, so that this model can be seen as a model that is applicable to humans as well as to other cognizing creatures, it could be said of this model that it is far too Hpo to be Hpl, because the same Hpo model could be applied to humans as well as to a homunculi-headed creatures, where a Hpl could not be applied to humans as well as to aliens.

### **3.3 ANIMAL AWARENESS**

This raises some important questions for comparative anatomy as concerned with questions about e.g. animal awareness and animal consciousness. Compare, for example, Gerald Edelman's thinking about how to determine whether animals are conscious, self-conscious, have qualia etc, of which he speaks in a section appropriately titled 'Problems of report in humans and animals'. He states that humans are in the privileged position of being able to report verbally on their behavior, reasons, intentions, and thus are the 'canonical referent' for scientific investigation into the questions mentioned. Questions about animal consciousness can only be inferred by '(...) behavioral, evolutionary, and morphological evidence.' (1989: 22-23) His suggestion is that we compare phenotypic and neural counterparts in humans and other animals, so that, if we:

(...) could determine the anatomical basis and functional role of consciousness in humans, and could demonstrate that animals possess similar structures and functions, this would provide additional grounds for the belief that they, [animals - MG] too, are conscious.

(1989: 22-23)

Edelman's search strategy seems, *prima facie*, fair enough and scientifically speaking very reasonable, so let us take it as our point of reference for the following.

If we would like to answer questions concerning animal consciousness we are bound to forms of comparison (of e.g. neurology, anatomy, functions). But to provide a measure for the comparison, models that are appropriate for this goal are needed. An Ha-model (or: an ideally complete description of a mechanism) would obviously be far too specific to be able to specify points that could be used for comparison, where a Hpo-model would be too general, so that it would include, as in the case of the homunculi-headed creature included in the range of mental creatures by Functionalism above, too little to capture any relevant points of comparison. What is needed here is something like a Hpl or a Ha model, but abstract enough to allow for comparison, in other words, a mechanism schema, that has abstracted away from the 'gory details', but that can, in the case of humans, be completely filled in so that no filler terms are used.

But, when a model is made with the explicit goal of using it for the comparison of human with *animal* cognition or consciousness, filler terms are necessary to give the model its intended liberalism. For if an animal were to use a slightly different mechanism than humans do (i.e. if there were a difference in *entities*, for a difference in *activity* would establish grounds for asking new questions), this would be a discovery worth noticing, not a reason to state that animals do not possess cognitive mechanisms or conscious states.

At the same time, the criteria that Craver (2006) (see section 2.4 above) sketched to distinguish Ha from Hpl models, can here be used in the opposite way, namely as a benchmark for keeping the model liberal enough. That would mean that, while comparing a mechanism schema of a certain human cognitive ability with animal anatomy, researchers should be careful that the schema they use is not (1) neglecting the different features associated with the phenomenon, because the cognitive ability under research could have different features in different species.<sup>23</sup> Next, they should pay attention to the possible different (2) precipitating, and (3) modulating conditions, because in the case of precipitating conditions, the same function can be dealt with

<sup>23</sup> This benchmark can also be grounded in biological theory, for, as François Jacob has emphasized '(...) evolution does not produce innovations from scratch. In works on what already exists, either transforming a system to give it a new function, or combining several systems to produce a more complex one.' (1982) Or, in other words, evolution is a 'tinkerer'. And as there is an evolutionary gap of approximately seven million years since the human branch split from the other great apes, some differences are bound to having been formed.

through different mechanisms, and in the case of the modulating conditions, variations in background conditions (which should not rarely be the case in different species), change the way the mechanism works. Also, the (4) byproducts can deliver valuable information, for although the byproducts might not matter to the phenomenon in question, they could provide researchers with important clues. For example, if an animal of which we want to know whether it has the cognitive ability  $A$ , has a mechanism to the effect of  $A$ , but that mechanism is more than slightly different so that the researchers do not immediately recognize it as a mechanism for  $A$ , the byproducts can point them in the direction of the view that the mechanism they look for is there, although performing in a different way than in humans. So the byproduct(s) might be similar where the actual mechanism is different.

The discussion so far also requires a skeptical remark. Because different mechanical ways can lead to identical results, a difficult, possibly unresolvable situation arises if we take Edelman's strategy as discussed above, into account. It could very well be possible that scientists will be unable to find out if animals have the cognitive features  $x, y, z$ , of which we know humans possess them, because if animals have another mechanism to the same effect, it may, without the verbal feedback humans can typically deliver about their behavior, reasons, intentions, be impossible to find out the function of the mechanism under research.

For example, imagine we knew that humans who claim that they lost their self-control while, for instance, attacking someone, spoke the truth because, using brain-scanners, we can see that neurotransmitter  $N$  blocks the part of the brain  $B$  associated with conscious control, and we have done experiments in which we artificially block  $B$  while the subjects are unaware of our doing so, and they also claim to have lost their ability for self-control. If we then wanted to find out if cats can also lose their self-control, and in cats, incidentally, instead of  $T$ , brain part  $C$  blocks the cats' brain part  $B$  (which can be described with a mechanism schema  $S$  which is also applicable to humans) associated with conscious control, the scientists investigating the question would be at a 'comparative anatomy' loss. This because they cannot use the argument of similarity. So, the mechanism approach also has a catch.

## IV | THE CHINESE ROOM, THE MECHANISM APPROACH, AND COMPUTER CONSCIOUSNESS

Another classic text in the philosophy of mind is J.R. Searle's 'Minds, brains and programs' (1982).<sup>24</sup> It builds further on Block's (1978) (cf. Block (1995: 267)) and is meant as a definitive argument against what Searle dubbed 'Strong Artificial Intelligence' (Strong AI), the current within artificial intelligence that holds to the thesis that an appropriately programmed computer literally *is* a mind (Searle 1982: 253). The first section of this chapter will feature an outline of Searle's (1982), in the second section I will cover new ground by applying the mechanism approach to the challenges introduced by Searle, after which I will propose the Principle of Double Isomorphism (PDI). Finally, in the third section I will discuss the implications of the PDI for the question of computer consciousness.<sup>25</sup>

### 4.1 SEARLE'S CHINESE ROOM

The weight of Searle's argument hinges on a thought-experiment commonly referred to as the *Chinese Room*. Searle asks us to imagine that he is locked in a room with a large batch of Chinese writings, and that he himself knows no Chinese. He is then given a second batch of writings and an English instruction manual on how to correlate the second batch with the first batch. He is then given a third batch of Chinese texts, again together with some English instructions that enable Searle to correlate the third batch with the first two batches. This last set of instructions also tells him how to respond, in Chinese writing, to the third batch. Unbeknown to Searle, the first batch of Chinese writings is a 'script', while the second batch is a 'story', and the third batch are 'questions'. The symbols he uses to respond to the third batch (the questions), are called 'answers to the questions', and the last set of instructions he got is called the 'program'.<sup>26</sup> As Searle gets used to the batches of Chinese text and the English instruction manuals, he gets so good that from the point of view of someone outside of the Chinese Room (the room Searle is working in), Searle's:

<sup>24</sup> Originally published in 1980, all references are to the reprint in Hofstadter & Dennett (1982).

<sup>25</sup> The term 'computer' in 'computer consciousness' is meant in a broad sense which also capture notions like 'machines consciousness' or 'robot consciousness'.

<sup>26</sup> In the following, I will use the words 'program' and 'system' interchangeably.

(...) answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody looking at [Searle's] answers can tell that [he doesn't] speak a word of Chinese. (1982: 355)<sup>27</sup>

Important here is that Searle emphasizes that he does not understand a word of Chinese, and thus is totally unaware of the functional role he is playing in the Chinese Room. Sitting in the Chinese Room, Searle is supposed to correspond to the hypothetical computer program written to understand stories and to give (correct) answers to random questions about the story. The room itself should correspond to the computer on which the program is installed.

Searle then uses the thought-experiment to try to clarify the claims of Strong AI. Strong AI claims that (1) a program written to understand stories and to give (correct) answers to random questions about the story does indeed understand the stories, and also claims that (2) the program explains human understanding. Searle disagrees with both these claims. He thinks that even if a computer program was able to achieve the results mentioned, it would not understand the stories, and neither would AI, by writing the program, explain anything about human understanding.

In applying the thought experiment to the claims of Strong AI, Searle reaches the following conclusions. In the case of the claim about understanding, it seems to him that as Searle himself, working in the Chinese Room, does not understand a word of Chinese, or, for that matter, anything about what he is doing, neither would the proposed AI program understand anything of the stories, since '(...) in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing.' (1982: 356) In the case of the claim about the explanatory value of the story-understanding program, Searle states that since the computer and the program are both functioning while lacking understanding, the computer and the program on it do not provide sufficient conditions for understanding. The supporters of Strong AI claim that what Searle does when he uses his understanding of English to answer English questions about English stories using English sentences, is formal symbol manipulation, and it is formal symbol manipulation when a computer program does the same thing (this thus constitutes a claim of isomorphism). About that, Searle says some things that are worth quoting in length, because they are necessary for a correct understanding of what he is claiming:

One of the claims made by the supporters of Strong AI is that when I understand a story in English, what I am doing is exactly the same -or perhaps more of the same- as what I was

---

<sup>27</sup> Obviously, this is a direct reference to the Turing Test, see: Turing (1950).

doing in manipulating the Chinese symbols. It is simply more formal symbol manipulation that distinguishes the case in English, where I do understand, from the case in Chinese, where I don't. (...) Such plausibility as the claim has derives from the supposition that we can construct a program that will have the same inputs and outputs as native speakers, and in addition we assume that speakers have some level of description where they are also instantiations of a program. On the basis of these two assumptions we assume that even if [the program] isn't the whole story about understanding, it may be part of the story. (1982: 356)

Searle then turns to six objections to his thought experiment and its purported conclusions: (1) the Systems-Reply, (2) the Robot Reply, (3) the Brain Simulator reply, (4) the Combination Reply, (5) the Other Minds Reply, and (6) the Many mansions Reply. Of these replies, the first, the third and the fourth are important in the context of this chapter.

The Systems Reply holds that even though it is true that the person inside of the room does not understand Chinese, the person is part of a larger system, which does, as a whole, understand Chinese. Understanding is thus, according to the Systems Reply, not attributed to the individual but to the whole system of which the individual is a part. To this, Searle's responds with an adaptation of his original Chinese Room thought experiment, in which we are asked to imagine that the man in the room memorizes the three batches and the two instruction manuals and does everything in his head instead of by hand. Searle then states that this would make no difference, for even the person who memorized everything will still not understand a word of Chinese, and thus, neither will the system as a whole. According to Searle, digressing on the status of AI as a branch of psychology, AI needs to be able to tell the difference between:

(...) systems that are genuinely mental from those that are not [and] (...) to distinguish the principles on which the mind works from those on which nonmental systems work; otherwise it will offer us no explanation of what is specifically mental about the mental. (1982: 361)

Searle then turns to the Brain Simulation reply, which states that if a system is designed that simulates all of the actual neuronal patterns working in the human brain of someone understanding Chinese, it will truly understand the story. Searle's answer to this Objection is that if a mechanism is constructed in which a man operated on pipes and valves which correspond to the actual neuronal firing patterns of a human Chinese-understanding brain, this machine would obviously not

understand anything about Chinese, thus, even a brain simulation would not understand Chinese.

The fourth reply, The Combination Reply, states that if someone builds a robot containing a program which simulates all the neuronal activity in the human brain, with behavior that is indistinguishable from human behavior, and a unified system, then, intentionality can be ascribed to this robot. Searle disagrees with this reply because he thinks that as soon as we know a formal program in virtue of which a subject understands and acts etc, is being instantiated in a being, we would not want to ascribe intentionality to it. (1982: 365-366)

For the purposes of the next section, it is also necessary to take a short look at the Reflection D. R. Hofstadter and D.C. Dennett wrote on Searle's 'Minds, brains, and programs' (Hofstadter and Dennett 1982). They point out what they think is the 'serious and fundamental misrepresentation' (Hofstadter and Dennett 1982: 373) that Searle lures us into, and which convinces us his thought experiment has the consequences he says it has. This misrepresentation consists in the fact that Searle convinces us that a human being could possibly do the things Searle imagines himself doing in the Chinese room. But this constitutes, as they say, '(...) an impossibly unrealistic concept of the relation between intelligence and symbol manipulation.' (1982: 373) Here, there is an enormous difference in complexity between two conceptual levels. At first, Searle invites us to imagine him hand-simulating an AI program, but the work that would actually consist in would progress at a painstakingly slow pace. Searle then, almost unnoticeably, switches the attention of the readers to a program that could pass the Turing test. Hofstadter and Dennett then argue for the Systems Reply; that it is a mistake to argue that because a part of the system does not understand Chinese, neither can the whole system.

## 4.2 MECHANISTIC LESSONS FROM THE CHINESE ROOM

As we have seen above, Searle disagrees with the two claims advanced by Strong AI, i.e. that (1) the program under discussion would understand stories and (2) that the program in question would explain understanding. We will first discuss Searle's analysis of the second claim, next we will look at Searle's criteria for isomorphism to establish a new principle concerning isomorphism of mechanisms and predicate ascriptions, and then we will discuss the impact that principle has on the Replies to the Chinese Room as discussed above.

As concerns the claim about explanatory value, the mechanism approach would here agree with Searle's conclusion, albeit on different grounds. Searle claims that because the AI program is functioning while lacking understanding, the program is not showing, and thereby not explaining, the sufficient conditions for understanding.<sup>28</sup> The claim of Strong AI depends on an assumed

<sup>28</sup> Note that this is a lesser claim than the claim advanced earlier: that the program would show understanding

isomorphism between formal symbol manipulation in computer programs and formal symbol manipulation in humans,<sup>29</sup> which Searle dislikes. As quoted above:

Such plausibility as the claim has derives from the supposition that we can construct a program that will have the same inputs and outputs as native speakers, and in addition we assume that speakers have some level of description where they are also instantiations of a program. (1982: 356)

On the mechanism approach, it is indeed the case that the description of a mechanism is an idealized description of the set-up and termination conditions, and of course there is a way to describe human understanding with the help of an abstract description. The question is whether that description is a Hpo, Hpl, or Ha model. Recall, a mechanism is said to explain a phenomenon only when a number of criteria have been met. This already delivers great problems for Strong AI's claim to explanation. Even if one takes the intermediate conditions into account, the description of the activities that take the mechanism from the start-up to the termination conditions, which has to show how the '(...) actions at one stage affect and effect those at successive stages (...) ' (MDC: 12), it would already constitute the end of the explanatory value where the mechanism approach is concerned. If one takes brain neurology in account, it is clear that a comparison of the ways in which the stages in the brain affect and effect each other are very much different from the way the stages in a computer program do. So Searle is, on the mechanism approach, right to claim that Strong AI's claim to explanation cannot stand its ground.

Searle's strategy appears to be one of finding isomorphisms between human and computer systems, as already becomes clear in his digression on the status of AI as a branch of psychology where he writes about the supposition that '(...) speakers have some level of description where they are also instantiations of a program.' (1982: 356). If 'describable as a mechanism', or 'describable by a mechanism schema' is substituted for 'instantiations of a program', this would redefine Searle's strategy in terms of the mechanism approach. Searle is even more clear on his strategy when he states that:

[w]e can't make sense of the animal's behavior without the ascription of intentionality, and we can see that the beasts are made of similar stuff to ourselves (...) Given the coherence of

---

altogether. Searle is here implicitly downsizing the claim to a form where the program is still granted explanatory value if it just shows a sufficient condition for understanding, instead of wholesale understanding.

<sup>29</sup> Where a further (hidden) assumption is that humans cognition indeed is a form of symbol processing, which is not altogether that obvious.

the animal's behavior and the assumption of the same causal stuff underlying it, and that the mental states must be produced by mechanisms made out of the stuff that is like our stuff. (1982: 365)

Searle is thus listing three aspects he seeks in isomorphism, which allow one to ascribe mental attributes (i.e. in this case intentionality) to other creatures than humans. First, the necessity of the ascription for making sense of something, secondly, similar causal stuff, thirdly, underlying mechanisms made of stuff that is like our stuff. However, the mechanism approach, as we have seen above, remains silent about the 'right' materials required for the correct performance of a function, which means that this condition has to be dropped within the mechanism approach. Thus, two criteria remain; the necessity-ascription condition and the underlying mechanism-condition. Combining Searle's quest for isomorphism with the mechanism approach, we can advance the *Principle of Double Isomorphism* (or PDI for short) which can be defined as follows:

Two systems can be ascribed the same predicates if they have the same start-up and termination conditions, can be described with the same mechanism-schema, and fall within the boundaries for isomorphism-ascription set by the Craver. Thus, if isomorphism of mechanism is established, isomorphism of predicate ascriptions follows suit, hence the word 'double'.<sup>30</sup>

In the definition given, the 'boundaries set by Craver' mean Craver's statement that '[a]n activity that happens at the wrong time, that takes too long, or that unfolds too slowly for a given role cannot fill that role.' (2001: 62) These thresholds for describing two systems that are functionally equivalent as both being the carriers of the same predicates (e.g. 'having a mental life') are thus narrower than the ones defined by Searle, which only ask for similarities between ascription-necessity and underlying mechanisms.

A few of the consequences of this PDI are the following. First, the PDI would instantaneously wipe the Chinese Room off of the world chart of the mechanism approach, for as neurons transmit their signals with mind blowing speeds, Searle, working inside of the Chinese room, would never be able to attain such speeds, thus the Chinese Room could never correspond to a suitably fast computer program in the first place. Thus Searle's arguments against Strong AI would have to be defined on the basis of another thought experiment. This provides additional grounds for

---

<sup>30</sup> Of course, a correct principle for the ascription of isomorphism would involve much more technical language and specifications, but the purpose of this Principle is to see how far we can with this minimal, provisional Principle, using the concepts of the mechanism approach.

Hofstadter and Dennett's claim that Searle delivers a 'serious and fundamental misrepresentation.' (Hofstadter and Dennett 1982: 373)

Secondly, the Brain Simulation Reply needs to be taken more seriously in the light of the PDI. If the role function of the Simulation had the same speed and timing, and unfolded at the same speeds as the human brain does, it would be a ready candidate for the ascription of the same predicates as is the case with human cognition. On this account, Searle's objection to this reply seem too far-fetched. A mechanism in which a man manipulates pipes and valves would never get past the Craver-boundaries.

Third, the Combination Reply, which invoked a robot capable of interaction with the outside world, with a program that exactly simulated the human brain's activities, with behavior indistinguishable from humans, would certainly count as a good example of an artificial system that could readily be ascribed the same predicates as humans get. Searle's objection that as soon as we would find out about this robot's formal program, we would abandon the predicate ascription, holds no ground because, *ex hypothesi*, the robot proposed by the Combination Reply satisfies two of the three aspects Searle seeks in isomorphism, where the third aspect Searle stated can be disregarded on the mechanism approach. The Combination Reply also meets the requirements of the Craver boundaries, as it clearly states that the program simulates the human brain exactly, and has behavior that is indistinguishable from that of humans.

### 4.3 CONDITIONS FOR COMPUTER CONSCIOUSNESS

As we saw above (in section 3.3) on the issues of comparative anatomy concerning questions about animal awareness, the mechanism approach can be helpful when trying to look for similarities between human and animal cognitive functions. In the context of this chapter, which was concerned with artificial systems replicating human cognitive features, a similar tactic has been followed.

Edelman's search heuristics, which were followed in the search for similarities between humans and other animals, can be seen to share features similar to the PDI above. Both partake in the search for similarities; in the case of Edelman, similarities between humans and animals, and in the case of the PDI, similarities between humans and artificial constructs. In both cases the mechanism approach is employed to deliver concepts that can guide the search. The PDI was built on (implicit) suggestions made by Searle in his attack of Strong AI. It is thus fitting that it can be used as a sort of benchmark for testing computer consciousness.<sup>31</sup>

The reasoning behind this proposal should be made clear. "Officially", the mechanism

---

<sup>31</sup> Searle was, in the first place, talking about *understanding* by computer programs. But we will cover new ground from here on.

approach has (as of yet) nothing to say about the question of 'machine consciousness', to use the broadest term available. But it seems logical, considering Edelman's search heuristics, to apply similar heuristics when looking for machine consciousness. The rationale behind this is that if a good search strategy when looking for answers to questions about animal awareness follows certain criteria, an artificial construct which can be measured according to the same strategy, can also be given similar judgments. For example, if a mechanism M that is found in both humans and animals, is embedded in a similar mechanism hierarchy, in an artificial construct that has the same specifications, or, defined more narrowly, is describable using the same Hpl-model or mechanism-schema, that artificial construct, should, as a matter of consistency, be granted the same function as it was granted in humans and animals. And if it can be granted the same function it should also be describable with the same predicate ascriptions.

Of course this is a one way strategy. Where Searle engaged in his battle against the ultimate thesis of strong AI that a suitably programmed computer literally *is* a mind, this was a specific philosophical dispute which cannot be solved by the mechanism approach. It could very well be the case that the ultimate thesis of Strong AI can stand its ground under philosophical scrutiny, so that a man operating on tubes and valves can indeed be said to understand Chinese, but that is a different matter than the question when it can be said of a suitably built machine that it can be attributed the same predicates as humans and animals are attributed with. *That* question can be answered with something like a better adapted and more technically specific application of the PDI.

The question whether a computer can have consciousness, awareness, understanding, or be ascribed intentionality can have two at least two differently defended answers. One answer can be given after a philosophical debate on the meaning of the terms and the sufficient conditions for attributing those terms. The second way is by looking carefully at what mechanisms constitute human cognition, and then building isomorphically similar machines. Humans ascribe each other predicates like 'intentionality' and 'understanding'. Those ascriptions are based, amongst other things, on behavior and speech utterances which are only possible in virtue of what are now considered to be mechanisms in the human brain. An animal with the same mechanisms (i.e. with a mechanism that is describable with the same mechanism-schema or a suitably liberal Ha-model), embedded in similar mechanism-hierarchies, has the same mental faculties and can also be ascribed with 'intentionality' and 'understanding'-like predicates. So, if a machine is built using the specifications of the human / animal mechanism, thus a mechanism that is describable by the same mechanism-schema or a suitably liberal Ha-model, this machine also deserves the same predicates.<sup>32</sup>

---

32 Such theoretical attempts have indeed been made (e.g. see Dennett (1981), chapter 11: 'Why you can't make a

## CONCLUSION

As promised, this conclusion will briefly summarize the content of this thesis, and also sketch a few possibilities for further interactions between the mechanism approach and the philosophy of mind.

In the **General Introduction** it was stated that talk of mechanisms is not exactly ubiquitous in the philosophy of mind, and that this thesis will investigate whether or not the philosophy of mind is justified in that lack of interest. In other words, this thesis searched for fruitful possibilities of interaction between the mechanism approach and the philosophy of mind.

**Chapter I** introduced Descartes and La Mettrie as two early exponents of mechanistic thinking, but where Descartes postulated a soul for humans, La Mettrie held to the thesis that the soul was nothing more than the organization of the brain. We then saw that the philosophy of science used to think laws were explanatory, but that that changed when talk of mechanism was introduced. From then on, the new idea was that a phenomenon had not been explained until a mechanism was described. Thus, the use of talk of mechanisms that could already be found almost 400 years ago in Descartes and La Mettrie, has as of recently also been embraced by the philosophy of science. The short discussion of the use of intuition in philosophy that followed, concluded that the conformity of theories to our intuitions should be a guideline and not a dogma during research.

**Chapter II** was devoted to an examination as well as a clarification of the concepts that are frequented in the mechanism approach. Amongst other things, we saw that mechanisms are divided into entities and activities which performed a function. Interactions have to proceed in a regular fashion and have to have productive continuity. Proposed mechanisms can be divided into the more loosely formulated mechanism sketches, the more empirically informed mechanism schemata, and, the ultimate goal of science, ideally complete descriptions, the intelligibility of which can be emphasized by the use of the three-way division into Hpo, Hpl and Ha-models. Filler terms were

---

computer that feels pain'; Edelman (1992), chapter 19: 'Is it possible to construct a conscious artifact'; Metzinger (2009): chapter 7: 'Artificial ego machines'). The above constitutes a new defense or philosophical foundation for those attempts.

explained to be terms that fill gaps in descriptions of mechanism, without admitting to do so. We also took a look at Craver's six criteria for normatively assessing mechanistic explanations, which were also suited to differentiate between Ha and Hpl-models, and saw that scientists necessarily take certain levels of the mechanism they investigate for granted, which is called the bottoming out of the mechanism. Finally, attention was given to two challenges to the mechanism approach, leveled by Woodward, and Eckardt and Poland respectively.

In **Chapter III**, I gave my own analysis of a paper in the philosophy of mind in terms of the mechanism approach, namely the attack on functionalism by Ned Block. The paper in question leveled two objections to functionalism, that of liberalism, the attribution of mentality to systems which do not have mentality, and that of chauvinism, the denial of mentality to systems that have mentality. Using the mechanism approach to analyze the debate, I showed that Block's notion of liberalism could be redescribed using the mechanism approach's division between Hpo, Hpl and Ha-models.

In the third section I showed that the difficulties with the ascription of mentality or nonmentality to homunculi-headed systems, elementary-particle-people and aliens as described by Block, pertained to the debate about animal awareness in the philosophy of mind. Using Edelman's search strategy, which tries to infer animal awareness by first establishing links between human verbal reports and anatomical structures, and then looking for similar structures in animals, I argued that that search strategy needed to be liberal in the description of neural mechanisms to avoid the problem of chauvinism. Next, I argued that filler terms are also necessary in that strategy, because it can very well be the case that animals employ different mechanisms than those possessed by humans, to perform similar functions.

The chapter was concluded with a skeptical remark. When looking for neural structures in animals that are similar to those found in humans, of which we know the function because of the availability of verbal feedback by humans, it could be the case that when an animal species uses a different mechanism, it cannot be established if that mechanism performs the same function as found in humans, because animals cannot give verbal feedback.

In **chapter IV** I offered a second investigation of an issue in the philosophy of mind, namely that of Searle's Chinese Room. Searle uses his Chinese Room thought experiment, in which he, locked in the Room, corresponds to the program hypothesized by Strong AI, to attack Strong AI which holds to the theses that an appropriately written program has understanding as humans do, and explains human understanding. Searle also defends his attack against certain replies, i.e. the Systems Reply,

the Brain Simulation Reply, and the Combination Reply. In applying the mechanism approach to that discussion, I argued that Searle, in his attack on the explanatory value of AI, was right for a different reason than the one he gave, i.e. that because computer programs are too different from the workings of the human brain to be describable with the same Hpo or Hpl-model as the human brain is. Then I used Searle's criteria for isomorphism to propose the Principle of Double Isomorphism (PDI), which states that if two systems are isomorphical under constraints taken from the mechanism approach, isomorphism of predicates should follow. Using the PDI, I then argued that Searle's use of the Chinese Room to attack Strong AI is untenable, that Searle's dismissal of the Brain Simulation reply was too far-fetched, and that the proposal of an advanced robot as given in the Combination Reply is an adequate reply to the Chinese Room objection. In the third section I argued that the PDI can be used as a benchmark to test computer consciousness, which means that if a robot is constructed that is isomorphical to the description of the mechanisms found in human brains, that robot should be ascribed the same mental predicates as humans are ascribed, as stated by the PDI.

That is, briefly, what I have done in the above. As we have seen in the sections 3.2, 3.3, 4.2 and 4.3 of this thesis, the application of the mechanism approach to the philosophy of mind yields interesting answers to old questions, interesting new questions, and many other possibilities for future research. Thus, to summarize the results of this thesis in one sentence: Cross fertilization between the mechanism approach and the philosophy of mind is possible as well as fruitful. This of course calls for new avenues for research. To name a few questions (in no particular order) that forced themselves to me while thinking about this thesis and writing it:

1. As to the history of philosophy, we saw in section 3.2 that Block had already made remarks about what the mechanism approach calls bottoming out. Are there other historical precursors to the terminology used in the mechanism approach?
2. According to Von Eckardt and Poland, externalistic theories of representation cannot be captured in the mechanism approach, but the mechanism approach allows for mechanisms to be embedded in hierarchies. What are the prospects for embedding human beings as mechanisms into the outside world? This seems a particularly pregnant question if one takes recent thinking about the mind as being extended into account (cf. Clark 2001).

3. The accurate description of a mechanism is said to explain a given phenomenon, but to what extent are mechanisms able to explain what e.g. 'pain', 'feeling', or 'suffering', consist in?
4. In the same vein, Paul M. Churchland's 'eliminative materialism' states that our commonsense conceptions of psychological phenomena will not stand up to a completed neuroscience, and have to be abandoned. What, given the mechanism approach, are the prospects for that prediction?
5. As we saw in section 2.4, Craver has sketched some criteria useful for distinguishing Ha from Hpl models. Is his list complete, and if not, what needs to be added to it or changed about it?
6. In general, as we have seen that the application of the mechanism approach to the topics discussed by Block and Searle have led to interesting results, what would the results be if the mechanism approach was applied to other topics in the philosophy of mind?
7. As has been emphasized in note 27, the PDI as defined above is just provisional. So, how would a sufficiently technically accurate Principle have to be formulated?

---

Closing this paper, I would like to repeat the quote from MDC, given at the very beginning of this thesis, which stated: 'Thinking about mechanisms presages new ways to handle some important philosophical concepts and problems.'

I hope the reader is, at least for the time being, convinced of its truth. As a final remark, I would like to state that, while looking back at this thesis, its claims, and its proposals, I sincerely hope philosophers of mind, now and in the future, will join me in appreciating the value of the mechanism approach, and join in to explore its consequences and applications, which seems, as I hope this thesis has made clear, very promising indeed.

## REFERENCES

- Bechtel, W. (in press) 'Constructing a philosophy of science of cognitive science', forthcoming in *Topics in cognitive science*.
- Bechtel, W. and Herschbach, M. (in press) 'Philosophy of the cognitive sciences', to appear in: Fritz Allhof (ed.), *Philosophy of the special sciences*. New York: SUNY Press.
- Block, N. (1978) 'Troubles with functionalism', in: Block N. (ed.) (1980) *Readings in Philosophy of Psychology*, Volume I. London: Methuen, pp. 268-305.
- Block, N. (2004) 'The mind as software in the brain' [edited extract from the article with the same name, for references see Heil (2004) p. 267], chapter 17, pp. 267-271 in: Heil J. (ed.) (2004) *Philosophy of mind: A guide and anthology*. Oxford: Oxford U.P.
- Craver, C.F. (2001) 'Role functions, mechanisms, and hierarchy' in: *Philosophy of Science*, Volume: 68, Issue: 1, pp. 53-74.
- Craver, C.F. (2006) 'When mechanistic models explain' in: *Synthese*, Volume: 153, Issue: 3 pp. 355-376.
- Cummins, R. (2000) "'How does it work" versus "What are the laws?": Two conceptions of psychological explanation', in F.C. Keil and R.A. Wilson (eds.) *Explanation and cognition*. Cambridge: MIT Press, pp. 117-144.
- Churchland, P.S. & Sejnowski, T.J. (1988) 'Perspectives on cognitive neuroscience' in: *Science*, Volume: 242, Issue: 4879, pp. 741-745.
- Clark, A. (2001) *Being there: Putting brain, body, and world together again*. Cambridge: MIT Press.
- Dennett, D.C. (1981) *Brainstorms: Philosophical essays on mind and psychology*. Cambridge: MIT Press.
- Dennett, D.C. (1989) *The intentional stance*. Cambridge: MIT Press.
- Dennett, D.C. (2009) 'The part of cognitive science that is philosophy' in: *Topics in cognitive science*, Volume: 1, Issue: 2, pp. 231-236.
- Descartes, R. (1998) *Discourse on method and Meditations on first philosophy*. (4th ed.) (Tr. D.A. Cress) Indianapolis / Cambridge: Hackett.
- Dretske, F. (1994) 'If you can't make one, you don't know how it works' in: *Midwest studies in philosophy*, Volume: 19, Issue: 1, pp. 468-481.
- Eckardt, B. von, & Poland, J.S. (2004) 'Mechanism and explanation in cognitive neuroscience' in: *Philosophy of science* Volume: 71, Issue: 5, pp. 972-984.
- Edelman, G. (1989) *The remembered present: A biological theory of consciousness*. New York: Basic Books.
- Glennan, S. (2002) 'Rethinking mechanistic explanation' in: *Philosophy of science* (Supplement), Volume: 69, Issue: 3, pp. 342-353.
- Hebb, D.O. (1949) *The organization of behavior: A neuropsychological theory*. New York: Wiley.
- Hempel, C.G. and Oppenheim, P. (1948) 'Studies in the logic of explanation', in: *Philosophy of science*, Volume: 15,

Issue: 2, pp. 135-175.

Hofstadter and Dennett (1982) 'Reflections', in Hofstadter D.R and Dennett D.C. (eds.) (1982) *The mind's I: Fantasies and reflections on soul and self*. New York: Bantam, pp. 373-382.

Jacob, F. (1982) *The possible and the actual*. New York: Pantheon Books.

La Mettrie, J.O. de (1996) *Machine man and other writings*. (Tr. & Ed. Thomson, A.) Cambridge: Cambridge U.P.

Machamer, P., Darden, L., and Craver, C.F. (2000) 'Thinking about mechanisms', in: *Philosophy of Science*, Volume: 67, Issue: 1, pp. 1-15.

Metzinger, T. (2009) *The ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Railton, P. (1978) 'A deductive-nomological model of probabilistic explanation' in: *Philosophy of science*, Volume: 45, Issue: 2, pp. 206-226.

Searle, J. R. (1980) 'Minds, brains and programs', *Behavioral and Brain Sciences*, Volume: 3, Issue: 3, pp. 417-458.  
Reprinted in: Hofstadter D.R and Dennett D.C. (1982), pp. 353-372.

Turing, A.M. (1950) 'Computing machinery and intelligence' in 'Mind', Volume: 59, Issue: 236, pp. 433-460.

Weinberg, R. (1985) 'The molecules of life', in: *Scientific American*, Volume: 253, Issue: 4, pp. 48-57.

Woodward, J. (2002) 'What is a mechanism? A counterfactual account', in: *Philosophy of Science*, Volume: 69, Issue: 3, pp. 366-377.

Wright, C. and Bechtel, W. (in press) 'Mechanisms and psychological explanation', to appear in Thagard, P. (ed.) *Philosophy of psychology and cognitive science* (Volume 4 of the Handbook of the philosophy of science). New York: Elsevier.