

Karlova Univerzita
Filosofická Fakulta
Ústav filosofie a religionistiky
obor: filosofie

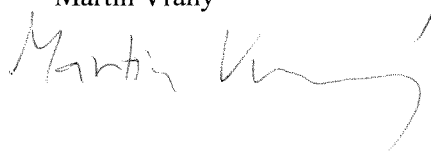
BAKALÁŘSKÁ PRÁCE

**D. C. Dennett's Approach:
On the Way to Explanation of Consciousness**

vypracoval: Martin Vraný
vedoucí práce: Prof. RNDr. Jaroslav Peregrin, CSc.
Praha, 2006

Prohlašuji, že jsem bakalářskou práci vykonal samostatně s využitím pouze citovaných pramenů a literatury.

Martin Vraný

A handwritten signature in black ink, appearing to read 'Martin Vraný', written in a cursive style.

Contents

1	Introduction	5
1.1	Preliminary Notes	8
1.2	Outline	11
1.3	Nature of the Explanation	12
2	The Method of Heterophenomenology	14
2.1	Facing the Explanatory Gap	14
2.2	Heterophenomenology	15
2.3	Intentional Stance	16
2.4	Heterophenomenological Worlds	19
2.5	Heterophenomenological Reduction	20
2.6	Incorrigibility of the First Person	22
3	Replacing the Cartesian Theater	24
3.1	Rejecting the Cartesian Theater	25
3.2	Introduction of the Multiple Drafts Model	28
3.3	The Multiple Drafts Model	29
3.4	Continuity of (the notion of) Consciousness	33
3.5	Making Contents Conscious	35
4	The Evolutionary Myth	39
5	Memes and Virtual Machines	44
5.1	Memetics	44
5.2	Mind as a Program	46
6	Semantics in the Brain	50
6.1	Searle and Semantics in the Brain	52
6.2	Understanding Memes or Memes Understanding?	53
6.3	Representation in the Brain	55
6.4	Dennett and the Chinese Room	58
7	Conclusion	60
	References	62

1 Introduction

Mind-Body problem has been a perpetual philosophical issue ever since the dawn of science, and the discussion of Descartes's dualism was the paradigmatic showroom of possible solutions for a long time. It wasn't perhaps until the mid-20th century that the debate underwent a considerable conceptual transformation thanks to two impetuses from science: the boom in brain-scanning experiments and the boom in computing devices and informatics. The former provided philosophers and scientists with abundance of evidence of correlations between brain activity and mental states. The latter showed how higher mental activities, like pattern recognition or playing chess, could be performed by fast computing machines running a relatively simple programme. Besides these two, there is another important source of influence (only as far as the methodology is concerned) – behaviourism. Despite its decline in 1960s marked by renewed interest in the study of mind, theorists were reluctant to rehabilitate the concept of consciousness. As Owen Flanagan puts it:

The irony is that the return of mind to psychology attending the demise of behaviorism and the rise of cognitivism did not mark the return of consciousness to the science of the mind. Mind without consciousness? How is that possible?

In the first place, the rejection of behaviorism did not take place with complete methodological abandon. A certain appropriate positivistic reserve remained.

...

In the second place, it seemed that one could map the mind, could provide a theory of intelligent mental life without committing oneself to any general view about the nature, function, or role of consciousness. [Flanagan, 1992, p. 3-4]

The basic principle of behaviourism, namely to take into account only the *observable* facts about one's intelligent and purposeful behaviour while remaining silent about mental states as a possible intermediary between causal input and behavioural output, seems to be rooted in the methodology of many theorists of mind, even though it is sometimes disguised under a less explicit form. D.C. Dennett is definitely one of them – he willingly admits to be a proponent of functionalism¹, at least as far as consciousness is concerned. His motivations for adopting functionalist stance may be various, but

¹As Dennett likes to put it, functionalism is the idea that handsome is as handsome does. More precisely, "functionalism is the doctrine that what makes something a thought,

the following surely plays a significant role: he tries to provide a *scientific* account of consciousness (that is an account which uses only verifiable claims or hypotheses expressed in terms that can be in principle part of the scientific discourse), so he has to accept, to a certain extent, scientific methodology; and as he points out, functionalism “is so ubiquitous in science that it is tantamount to a reigning presumption of all of science” [Dennett, 2005, p. 17].

If Dennett is such an adherent of scientific methodology, is his work still philosophically relevant? Most of the philosophers concerned with the mind-body problem, his many opponents notwithstanding, think so, for he not only spends considerable time demystifying the notion of consciousness (in which he seems to continue the work of Gilbert Ryle), but there also follow many philosophically interesting implications from his theory of consciousness. In order to understand how a scientifically oriented approach can give rise to philosophical questions, consider the classic Turing Test. Alan Turing, in his famous article *Computing Machinery and Intelligence* (1950), proposed to address the question “Can machines think?” by asking whether a machine could in principle pass a well-designed test. The test is an imitation game where a human interrogator is having a conversation with either a machine or another human being. If the machine can fool the interrogator, so that he thinks he is talking to human being, then the answer is yes, machines can think. Of course, the criteria of the test are blatantly behaviouristic, and the intuitive objection is: “But it doesn’t follow that the machine *really* thinks, it only produces certain speech acts *as if* it was conscious.” The intuitive appeal of this objection is powerful and the same rationale underlies another famous thought experiment – Searle’s Chinese Room argument against the prospect of the strong AI². Nevertheless, if we ask ourselves thanks to what we regard other people as conscious and rational, we will not end up with anything that could not be *simulated*³ by a machine. In other words, we

desire, pain (or any other type of mental state) depends not on its internal constitution, but solely on its function, or the role it plays, in the cognitive system of which it is a part” [Levin, 2004, section 1]. Therefore, as anything can rightly be called a heart, for example, if it serves the right function in the complex machinery of a living organism (namely to pump blood and thereby ensure its circulation all over the body), so the identity of a mental state can be said to be determined solely by its causal relations to sensory stimulations, other mental states, and behaviour.

²I expect the reader to be at least roughly familiar with the Chinese Room argument. A concise summary of Searle’s argument can be found in [Moural, 2003, pp. 216-226]; in this paper, I will consider the version of the argument stated in [Searle, 1990].

³It has been a perennial issue among the proponents of strong AI and their opponents whether a simulation of consciousness is actually its duplication. Obviously, a simulation of storm on a computer won’t make it rain in the lab, but as will be discussed later, Dennett

apply something like the Turing Test when we want to decide whether a being is intelligent (or conscious), though we do not do it consciously (or intentionally). This is not to show, of course, that Turing was right and that his criteria are sufficient, but rather that we should be careful not to trust our intuitions too hastily, for they may be deceptive. And one of the greatest assets of Dennett is his ability to expose many of our common intuitions about consciousness which he subsequently doubts and deems misleading or flawed.

Let's turn to the brain. Neurology and especially new brain-scanning techniques provided us with overwhelming evidence that the brain is *somehow* responsible for consciousness (mind). The core of the mind-body problem is the "somehow," and this paper will review Dennett's solution as expounded in his book *Consciousness Explained*. Before we start with Dennett himself, however, we can distinguish two metaphysical positions that describe the underlying ontological structure of the mind-body problem: dualism and materialism⁴. Descartes defended dualism, and despite its initial appeal, the problem of the causal interaction between the two substances seems to outweigh the benefits of the position (at least insofar as the mind-body problem is concerned). His followers amended his theory in an interesting way that led to occasionalism, but even though occasionalism is logically and metaphysically possible, it is not at all informative⁵. Materialism, on the other hand, faces equally serious difficulty: how could something like consciousness arise on the basis of a mass of unconscious neurons? Some philosophers (Nagel and Searle, for instance) are middle of the road, maintaining property dualism that claims there is only one substance (matter) that has yet two kinds of properties irreducible to one another – mental and physical. The substance dualism of Descartes seems to be definitely out of date nowadays. Therefore the debate on the ontology of the mental seems to be between property dualists and materialists. However, what many of them have in

holds that the mind is to the body as software to hardware – that there is nothing to the mind that could not be analysed in functional terms and hence embedded in a program.

⁴Logically, dualism is opposed to monism, not to materialism, but since materialism is a kind of substance monism, and the other possible kind of substance monism – idealism – seems to be completely abandoned nowadays, this opposition is usually accepted.

⁵Occasionalism is actually a token-token identity theory of mental states, and its close kin – parallelism – is perhaps not far from emergentism, one of many 'isms' that have arisen in the relation to the mind-body problem recently. It is not informative because it, by definition, does not explain the mind-body causation. Whether it is God or a pre-established harmony that makes up the illusion of real causation between mind and body, there is, in fact, no causation at all, and hence no scientific explanation is possible.

common is that they⁶ are naturalists: they think consciousness is a natural phenomenon produced by the process of natural evolution, at the beginning of which there were non-conscious organic molecules.

To make the starting situation clear, let's point out what is agreed on by almost all the theorists of consciousness. First of all, it is accepted that the brain (or rather the overall nervous system) is "the organ of consciousness," therefore the question is no longer whether the brain is after all (causally) responsible for consciousness, but *how* consciousness occurs upon brain processes⁷. Second, neurophysiological data are taken to be facts about brain processes and their temporal correlation with mental states is seldom disputed. Therefore, if there is a painful stimulus, the subject reports pain, and the neurophysiologist observes activation of a distinct area in the brain, we may rightly assume there is some kind of causal relation between the brain state and the mental state (though the metaphysical and physical nature of the relation remains unknown). Third, consciousness is taken to be a phenomenon which is in principle scientifically explainable and which should not be omitted if we are to explain human intelligence. This point is perhaps a matter of greater controversy than the previous two, but even eliminativists like the Churchlands address the concept of consciousness with respect. Last but not least, the epistemological problems of the existence of external world, possibility of knowledge etc. are put aside.

1.1 Preliminary Notes

Consciousness Explained is the name of Dennett's book whereby he expounds the fundamentals of his theory of mind. Although the title may seem to be very bold at the first sight (considering the number of theorists claiming that consciousness principally defies explanation), it should rather be understood as expressing Dennett's optimism as both to the possibility of scientific explanation of consciousness and the appropriateness of his own approach.

⁶To name only some of the prominent philosophers concerned in the debate: Searle, Nagel, Dennett, McGinn, Flanagan.

⁷To say "the brain is (causally) responsible for consciousness" is not to imply that consciousness is a real thing, a real link in the causal chain as known to physics. Beside the classic, "mechanistic" sense of the notion of causality, we can recognize a different sense – that in which some matter of fact (or structure) at a lower level of description is the cause of a more complex phenomenon at a higher level of description. In this sense, the liquidity of water is caused by its molecular structure. It is perhaps only a *quasicausation*, since it does not involve any temporal extent, unlike the "mechanistic" causation. Nevertheless, accepting this distinction allows us to speak about causal dependency of consciousness on the brain, no matter whether we consider a reductionist theory or any other.

The last part of the last chapter, titled *Consciousness explained, or explained away?*, shows that Dennett is well aware not only that his explanation is far from complete, but also that some may deem his attempt as no explanation at all, though for reasons which Dennett would never accept. What makes Dennett's project susceptible to principled objections is that he spends almost no time at all making the *explanandum* clear. The only part in the whole book where it is made clear what counts as the *explanandum* of his theory consists of two sentences:

In the chapters that follow, I will attempt to explain consciousness. More precisely, I will explain the various phenomena that compose what we call consciousness, showing how they are all physical effects of the brain's activities, how these activities evolved, and how they give rise to illusions about their own powers and properties. [Dennett, 1991, p. 16]

It is hard not to see Dennett's functionalist attitude which underlies the whole project. As Dennett often puts it, handsome is as handsome does, hence also consciousness is what consciousness does. Consciousness is, for Dennett, a set of phenomena, like intentionality, reflexion, experience, etc., which can be defined in functional terms – i.e. they can be characterized by their contribution to the cognitive machinery of human beings⁸. Some theorists, like Searle, Nagel or McGinn, may feel appalled by such a reduction, objecting that this already settles the whole issue. However, Dennett has at least two good reasons for such a reduction.

I have already mentioned the first – as he wants to do the science of consciousness, he accepts the scientific methodology. Science may never access some mysterious intrinsic properties that are postulated by laymen or simply “felt.” Hence, if science is to explain heat, it will concentrate on the explanation of what heat does (physically), such as melting, burning, warming etc. If the problem is settled this way, then to say that heat is a molecular motion is indeed the explanation, though we may thereby feel deprived of the idea why heat is felt as *heat*. Dennett is a verificationist, which is not surprising, given his functionalist attitude and often expressed sympathy for behaviourism⁹;

⁸This is, by the way, the basic reason for Dennett's rejection of the philosophical notion of *qualia*.

⁹What makes verificationism close to behaviourism is that following the basic tenet of the former, that a sentence is factual and meaningful only if it can be verified by empirical observation, results in a tacit motivation of the latter, namely that mental states *per se* should be left out since there is no *objective* way to observe them. Such a position, of course, does not deny the existence of mental states but it tries to avoid them in the explanation of human cognition.

he resists taking into account intrinsic properties of anything, consciousness notwithstanding. It is, of course, questionable whether it is still appropriate to maintain verificationist and functionalist position when it comes to consciousness. Dennett thinks it is, but perhaps not so much because he would insist there is no problem with such an approach, but rather because he realizes it may well be the only way to get started. At the end of *Consciousness Explained*, he remarks:

Only a theory that explained conscious events in terms of unconscious events could explain consciousness at all. If your model of how pain is a product of brain activity still has a box in it labeled “pain,” you haven’t yet begun to explain what pain is, and if your model of consciousness carries along nicely until the magic moment when you have to say “then a miracle occurs” you haven’t begun to explain what consciousness is. [Dennett, 1991, pp. 454-455]

This is the second reason, and it stems from what I call Dennett’s pragmatism. He often rejects some rebuttals of his arguments simply for the reason that the rebuttals and thereupon amended theories prescribe no research project. At the beginning of *Consciousness Explained*, he tries to show why he thinks dualism is forlorn and the decisive point is that dualism is antiscientific, not that it violates the physical principle of conservation of energy, Occam’s razor, or that it may never adequately explain the mind-body causation¹⁰. Dennett wants to escape from the seemingly neverending debate whether consciousness is in principle irreducible to physical properties or not simply by realizing that if we are to explain consciousness, we have to start with some presumption or other. What makes the debate about the irreducibility of consciousness in his eyes less fruitful than it seems to other theorists is his conviction that we already have sufficiently rich conceptual scheme and scientific framework to get down to the scientific explanation of the mind-body problem. The only thing we lack is, Dennett insists, a complex enough account of consciousness as arising on the brain processes. However questionable this Dennett’s conviction may be, it justifies his optimistic vision of the upcoming science of consciousness. And even if Dennett is wrong and a progress both in the conceptual and scientific framework must be made before we can successfully tackle the mind-body problem, Dennett’s work will make visible what concretely has to be done first and where the conceptual gaps are.

¹⁰Cf. [Dennett, 1991, p. 37]

1.2 Outline

The purpose of this work is not to address the problem of the irreducibility of consciousness, trying to weigh the pros and cons of both views, but rather to review Dennett's theory as a whole and to try to find that part of his theory which needs most further clarification in order for me, and all the people unable to fully grasp his theory, to recognize the theory as a full-fledged explanation of consciousness (and hence also the mind-body problem). This is not so easy as it may seem, for Dennett's theory is, as he himself readily admits, too counterintuitive to be easily comprehensible. It is counterintuitive in that it requires us to radically reconsider our usual way of thinking about consciousness. Not only does it undermine most of the beliefs and intuitions we have about consciousness, but it also robs us of that which we like most – qualia, the purported privileged access to our own mental states, and even the self. Since Dennett's theory poses many difficulties, I will be glad if I manage to interpret his theory without making an assertion in his name that would be contrary to what he has intended.

I will begin with *heterophenomenology* – a method of acquiring and interpreting the data about subject's mental states that conforms to scientific methodology. Next section will be devoted to Dennett's systematic rejection of the idea of Cartesian Theatre, which is a recurrent theme in his works and so will be in this paper. Then I will try to present the Multiple Drafts Model, which Dennett holds to be the right replacement of the flawed Cartesian Theater Model of consciousness. While the Multiple Drafts Model mainly tells us how consciousness is structured and related to brain processes, the story of the origin of consciousness in evolutionary terms is to show how such a complex phenomenon as consciousness can arise thanks to relatively simple processes and mechanisms; in other words, it fleshes out the Multiple Drafts Model that is first hard to grasp. At the end of the evolutionary story, there is the *homo sapiens* with his ability of speech, which enabled a new kind of evolution – the meme evolution – of which consciousness is a product, as Dennett claims. This, together with the well-known claim that minds are programs, will be discussed in the section "Memes and Virtual Machines." The last chapter will finally deal with what I think is the most unclear part of Dennett's explanation of consciousness, namely how mental states bear meanings. I will try to keep the reader alert to this problem throughout the paper by pointing out those parts of his theory which, I think, hinge on an account of meaning of mental states.

Generally, I assume that Dennett's approach to consciousness is meaningful even if we don't agree with some of his presumptions. For if we think of his project not as of a bold enterprise that insists to capture everything there

is to consciousness, but rather as of an attempt to explain as much as possible with the contemporary scientific knowledge, we can engage in the general endeavour for the scientific understanding of consciousness either by pointing out where Dennett explains something unsatisfactorily, or by showing what important feature was neglected.

1.3 Nature of the Explanation

Before I begin to scrutinise Dennett's theory, I have to address a general methodological problem: "What would count as an explanation of consciousness?" This question cannot be answered in terms of a clearly delineated *explanandum* to be accounted for by generally acknowledged concepts and principles, because there is no general agreement as to what consciousness is¹¹. However, the subjective experience of understanding or realization, that usually occurs when an explanation is successful, might provide the right clue for telling how a theory has dealt with its *explanandum*. For our purpose, I will first distinguish three kinds of realization, and then I will consider what kind of realization should occur if the explanation of consciousness is to be deemed successful.

"analytic" kind : A realization of the first kind happens as the outcome of apriori reasoning. Suppose a student knows both logarithmic and exponential functions and he can make calculations with them. If he hasn't realized himself already, somebody can explain him, or better say, show him that they are inversion functions. This is a realization of an analytical truth and it is accompanied by the well-known "aha" effect.

"empirical" kind A realization of the second kind has the form of an answer to a question like "Why is that so?" or "How does it work?" As an example, we may consider a youth pondering about how it comes that a car moves thanks to petrol. It may well seem miraculous at the first

¹¹Consciousness, in a sense, defies exhaustive description, and despite several attempts, there is no single and canonical definition of consciousness. Naturally, we can try to settle its distinctive features, but such a set of properties will always be prone to be challenged, since it may not capture some other important features, let alone the holistic account, which many theorists strive for. Nonetheless, we intuitively know what consciousness is and thence what is to be explained. Similar conceptual difficulties have accompanied the discussion about the concept of life: it is a fundamental question whether life should be described as a specific set of properties that form the necessary and sufficient conditions for something to be classified as living, or as a loose cluster of properties none of which is a necessary condition.

sight that a few liters of petrol make the car go for many kilometers, but if you come to know the mechanism of the car and the principles of the combustion engine, then you know “why,” and the potential of the car to move was explained. Surely, as the explanation avails of aposteriori reasoning, there remain several steps unexplained, such as “Why the mixture of petrol and air explodes when there is a spark (or high pressure)?”, but every aposteriori explanation must stop somewhere and the explosion is a familiar piece of everyday knowledge about the physical world. Again, there is the “aha” effect, despite taking some facts for granted.

“scientific” kind A realization of the third kind happens mostly after a scientific explanation that uses terms and “things” that are beyond our everyday experience. As an example illustrating this kind may serve the explanation of heat as a molecular motion, or falling of objects as the pull of gravity. There is no “aha” effect, and rather than a realization, it should be called an acceptance of a plausible model, for a layman would feel there is something fundamental missing - the explanation of why heat is felt as *heat* or what the gravity is. Unlike in the domain of macromechanics from the combustion engine example, here we do not know what terms like gravity or molecular motion factually refer to. Hence we don’t know whether gravity is *really* some force out there or whether we should regard it as an operational concept such as atom.

What kind of realization would we have to experience to judge some theory as an explanation of consciousness? Certainly not the “analytic” kind, for such a realization is impossible to occur, given that the explanation would depend on aposteriori reasoning about brain processes. If the explanation gave rise to the second kind of realization, we would find it satisfying, because we would thereby see how the brain produces consciousness, how it works. Such an explanation might entail some unsolved subproblems, similarly to the explosion problem in the combustion engine example, on the condition that these are potentially solvable by further progress in science and more importantly, that the absence of the solutions does not prevent us from understanding the overall theory¹². On the other hand, such an explanation would be most probably done in causal terms, hence consciousness would be

¹²Considering again the combustion engine example, the lack of knowledge about the explosion does not hinder us from understanding the whole mechanism - we just know that the mixture of petrol and air explodes. Some subproblems, however, may prove essential to the explanation of consciousness. This is the case, I think, of the recently proposed ideas from quantum physics, such as Penrose’s theory of the essential role of the quantum effects at microtubuli in neurons. The causal structure of quantum physics and its relation

treated as a causal product of brain processes, which is an option unavailable to reductionists who disregard the reality of consciousness, and who reduce it to some physical property of other.

So, whether such an explanation is even possible is an issue related to the so-called explanatory gap and to the problem of the irreducibility of consciousness. What can be settled for sure, however, is that Dennett's *Consciousness Explained* does not offer such an explanation. At the outset of the book, Dennett warns us that his theory is too counterintuitive to be understood after first reading, yet even if we read the book several times, we will not, I daresay, arrive at a realization of the "empirical" kind. However disappointing this so far unsupported claim of mine may seem, it does not follow that Dennett's theory is entirely forlorn, though it suggests that the title of the book promises more than can be found within. For even if *Consciousness Explained* offers "only" a theory leading to a realization of the "scientific" kind, it still contributes substantially, for it explains the structure of consciousness as functioning upon the brain and thereby it enhances our conceptual apparatus and proposes new research projects. In a sense, Dennett cannot offer an "empirical" explanation, since he has good reasons to believe that what we call consciousness is rather a fiction than a real thing – and hence cannot be explained as a causal product of brain processes. His explanatory strategy can be roughly summarized as follows: (1) he tries demonstrate that our knowledge of what consciousness is consists of many false and misleading intuitions; (2) he focuses on those features of consciousness which he finds functionally relevant; (3) he explains how the relevant features are physically realized in the brain; (4) finally, he shows why consciousness seems to us to be what it seems.

2 The Method of Heterophenomenology

2.1 Facing the Explanatory Gap

Dennett believes he is able to tackle seriously the question of the relation between the brain and consciousness, and if he is to succeed, he has to show how the explanatory gap between physical processes and consciousness can be overarched or dissolved.

What is the explanatory gap? It is a useful term, coined by Joseph Levine, referring to the apparently ultimate conceptual discrepancy between descriptions of the physical facts and descriptions of mental states as viewed from

to the macroworld needs to be clarified more in order for Penrose's theory to be generally comprehensible and informative.

the first person perspective¹³. By way of illustration, consider the classic example from neurophysiology: pain is C-fibre activation¹⁴. What does this simple assertion mean? Note that the claim is ascribed to a neurophysiologist, hence we should try for a charitable interpretation that doesn't take it too literally. Let's assume the neurophysiologist meant that what we refer to by the word "pain" in the utterance "I am feeling pain," is nothing but a pattern of activation of C-fibers. Furthermore, let's assume that the neurophysiologist has observed constant correlation between a subject reporting pain and an activation of C-fibers in his nervous system. The constancy of the correlation makes him think that the reported pain is *really* the activation of C-fibers and he therefore asserts the identity of the two. What kind of explanation is that? Does it explain what really matters about pain - its terror, unbearability, the "feel" of it? Obviously not. Isn't it what we would expect from every explanation of a mental state based on a physical state? What if we asked the neurophysiologist: "But why the C-fibre activation is felt like pain rather than like an itch or a tickle?" These problems aim at the conceptual disparateness of the science of the brain on the one hand and our phenomenology¹⁵ on the other, and they also seem to justify the use of the term "explanatory gap," in spite of the fact that some theorists, Dennett included, do not think there really is some explanatory gap.

2.2 Heterophenomenology

Now we have the idea of the explanatory gap, and we know what is to be done if we strive for a meaningful scientific explanation of mental states: a suitable conceptual framework must be devised. This is the purpose of Dennett's heterophenomenology¹⁶. Heterophenomenology is designed to be:

the *neutral* path leading from objective physical science and its

¹³The explanatory gap is sometimes referred to as the irreducibility of the phenomenological first person perspective to the scientific third person perspective, an issue first put forward by Thomas Nagel in his notorious article "What Is It Like to Be a Bat?" (see References).

¹⁴The matter is much more complicated nowadays, but since the identity of pain and C-fibre activation has become a paradigmatic example of reductive explanation, I let it that way.

¹⁵In the Anglo-Saxon philosophy of mind, the term "phenomenology" refers to the subjective character of our mental states, the what-it-is-likeness of, for instance, feeling pain, seeing red, smelling a perfume, feeling desperate etc. The name of the philosophy of Husserl and his disciples is capitalized: the Phenomenology.

¹⁶Dennett himself points out that heterophenomenology is a common practice in the science studying human and animal consciousness; he insists that he only describes the method and shows its rationale. See [Dennett, 2001, section 1.] or [Dennett, 2005, p. 36]

insistence on the third-person point of view, to a method of phenomenological description that can (in principle) do justice to the most private and ineffable subjective experiences, while never abandoning the methodological principles of science. [Dennett, 1991, p. 72]

This is still too abstract to see what exactly is needed. The way the science of consciousness makes progress is by gathering data from experiments and interpreting them. Subjects of the experiments are always conscious beings (mostly adult humans) and there are two main kinds of data: (1) those observed by various measuring devices (brain scanners, eye trackers, stopwatch etc.), hence third-personal and “objective,” (2) reports of the subject on his or her occurring mental states, hence first-personal and exhibiting intentionality. How to approach the latter kind of data in a way that would be compatible with the way we deal with the former kind of data?

Dennett proposes first to record all the subject’s utterances in the objective way as a tape-recorder does – these are the raw data for heterophenomenologists. Note that the data bear no meaning at this stage; they are just recorded noises yet to be interpreted. The second step is crucial and nontrivial, even though it is a common practice: we adopt the intentional stance and thereby we interpret subject’s utterances as speech acts. Here I have to make a digression to explain the notion of the intentional stance.

2.3 Intentional Stance

The intentional stance is a descriptive and projective stance adopted by an observer who, by observing certain patterns of behavioural reactions of the system, understands the system as rational. The intentional stance is descriptive in the sense that the observed behaviour is described in intentional terms (e.g.: “The bird wants to be fed.”). It is projective in the sense that intentionality is in the eye of the beholder. In order to clarify the nature of intentional states according to Dennett, let me quote J. Haugeland’s concise explanation:

A system has intentional states - paradigmatically, beliefs and desires - just in case its behavior exhibits a specific sort of observable pattern. The intentional states aren’t the same as that observable pattern, or, at least, not the observable part of it; rather they are a kind of completion of the pattern that is more or less necessary for it to stand out clearly in the first place. So it’s roughly as if you were given every other letter of a text, or

a scattered fraction of an image, and you “fill in” the “missing” pieces. Without that filling in, the visible part seems irregular and disjointed; yet its structure becomes conspicuous and compelling, once the remainder is interpolated.

...
Intentional states and processes are, in principle, *nothing but* our projected filling in of the pattern, in such a way that it makes sense overall. This is the way in which they are “in the eye of the beholder” - or perhaps it would be better to say, in the sense of the understander. But note well: this by no means renders them fictional, gratuitous, or arbitrary. In any given case, the attribution of intentional states is strongly constrained, both by principles and by facts - so strongly, indeed, that it is a nontrivial *achievement* to succeed at it at all. [Haugeland, 1998, p. 292]

Thus the intentional stance provides the observer with a stable and meaningful interpretation of the behaviour of a system. For instance, if we observe a rotating mill-wheel, we don't need to ascribe intentions to it, because we can interpret its behaviour in much more economic and stable way: it is the water flowing beneath that makes it rotate. On the other hand, if we observe a bee flying around and sitting on blossoms, we can't even help it not to interpret the bee's behaviour as motivated by intentions. Filling in the intentions is postulating the intermediary between the system's behaviour and its environment or close surroundings (which includes all the stimuli that come from the environment). This intermediary is sometimes necessary for a stable and meaningful interpretation - we can do without it in such simple cases as the mill-wheel, but we apparently cannot do without it when interpreting others' behaviour, for example. I think Dennett suggests that it is rather a matter of complexity of the observed behaviour what makes us adopt the intentional stance, not some intrinsic property of the system (like *intrinsic*, innate intentionality, as Searle believes). Therefore we may justly speak about a robot's intentions, or an anthill's¹⁷ intentions, as well as about human's intentions. Moreover, according to Dennett, robots and anthills can be said to harbour intentions in the same way as we do - they just may not be conscious of it! The complexity of behaviour of higher species and some artifacts such as robots or computers is so immense that we help ourselves by clustering pieces of information into greater wholes, we take certain pattern of a system's behaviour to be an expression of an intention. For example, suppose we observe a bee moving confusedly in the vicinity of its fellow bees

¹⁷For a description of an anthill as a whole exhibiting intentional behaviour see D. Hofstadter's eye-opening book *Gödel, Escher, Bach*, chpt. *Prelude... ant Fugue*.

that subsequently fly away right to the place where the first bee discovered a nice blossom. Adopting the intentional stance then, we interpret the dance of the bee as telling the way to the other bees, paying no extra attention to the details of the dance and the exact system of derivation of the direction out of the apparently confused motion. If we knew more about bees and their internal constitution, however, we could understand the link between the bee's discovery of a new blossom and its dance without the intermediary of intentions. Consider once more a fictive case of an extraterrestrial observing a mill-wheel: it might think "The wheel turns around because it wants to crush grain via the millstone attached to it." It is a stable interpretation, though not at all rational from our point of view, and several other problems remain (such as: "Why would the mill-wheel want to crush grain?"). Moreover, ascribing intentions to something makes us see the analysed performance as a *purposeful* action. We do not ascribe intentions to water to flow down whatever the cost (since we can explain this action by laws of physics), hence we do not regard it as a purposeful action. On the other hand, whenever we ascribe intentions to something, we thereby understand it as acting for purpose. It is a trivial statement (insofar as "purposeful" and "intentional" are synonyms), but if we knew more about the inner constitution of, say, a bee, we would understand its behaviour (actions) as a physically grounded effect. The purposes may still be there, however, hard-wired in the inner constitution by the process of evolution. They just may be out there, without anyone appreciating their guiding role. This will be discussed properly later in the section on the Evolutionary Myth.

We see then that Dennett's notion of intentionality does not presuppose consciousness, at least not at the side of the "intention-possessor." We should keep in mind, however, that when interpreting something as intentional system, we thereby situate a self of the system. Clearly, every intention must be intended by something. Such a posited self is as fictional as the intentions of it, it is only an abstract center to which all the intentions must be bound. Dennett names it *the center of narrative gravity* and he intends this notion to replace the traditional "realistic" notion of the self. I hope to address this issue later; for the time being it is only important to see that Dennett's early theory of the intentional stance underlies many of his subsequent speculations.

What about the intentional stance itself? Is not the ability to adopt the intentional stance conditioned by the faculty of understanding one's own intentions? Do we not interpret others' behaviour as intentional because we ourselves harbour similar intentions? Do animals adopt the intentional stance too? Imagine a cat flees away as you swiftly raise your hand. Did it run away because it ascribed to you (or to the hand) the intention to hit it? If

such an explanation is meaningful, what would such an intention-ascription be like? It is nothing more, I guess, than a sort of expectation. It is indeed a possible *functional* explanation of intention-ascription: we do it in order to be able to predict others' behaviour, which is obviously advantageous (hence justified from the evolutionary point of view). But does it capture all there is to understanding one's intentions? If understanding what others do, and why, consists solely in ascribing intention to them, which itself is a procedure devised mainly to enable successful predictions, then the difference between human and animal understanding of behaviour is only a matter of complexity. This is not to suggest there is something wrong with such a view, it only seems to overlook those cases when the understanding largely depends on one's familiarity with a similar action. If you see a child joyfully running back and forth, you will perhaps understand its behaviour on the basis of being familiar with a similar state of mind from your own childhood, rather than thanks to seeing a purpose in that action. This case might be misleading, I admit, for it is disputable whether such a behaviour is intentional. I only want to pay attention to the intuition that we understand what others do thanks to some kind of acquaintance with our own intentions¹⁸.

2.4 Heterophenomenological Worlds

Having an inkling of what the intentional stance is, we may continue with heterophenomenology. By adopting the intentional stance, we interpret subject's utterances, button-pushes¹⁹ and other relevant results of the experimental task as speech acts expressing subject's beliefs and desires.

No matter what the nature of our conscious beliefs is, it is important to isolate subject's reports on his own conscious mental states from other signs of his mental activity which can be measured objectively and without subject's contribution, such as galvanic skin response, eye movements, EEG, etc. Upon these subjective reports, the experimenter can reconstruct subject's heterophenomenological world, that is the world of subject's beliefs, desires etc. as inferred from the set of all the speech acts performed during the experiment. Dennett likens this step to the process of reconstructing the

¹⁸Besides, the core of the so called "argument from analogy" for other minds states that our knowledge of the minds of others is based on extrapolation of workings of my mind to other people. I leave to the reader whether this remark is relevant or not, for Dennett might have just intended the intentional stance to replace this mind-to-mind interpretation of others' behaviour. On the other hand, he himself introduces a similar, if not the same, explanation of how we understand others - see p. 21.

¹⁹Pressing a button is often used as a shortcut for production of a specific speech act, for instance "My interpretation of the Necker Cube has just changed."

fictional world of a novel: regarding the text of a novel, we can meaningfully speak about what is true in that fictional world (that Desdemona was faithful to Othello, for example), and more important, we can treat it with the required objectivity, because these fictional facts are intersubjectively confirmable (every reader of Othello would confirm that Desdemona remained faithful).

2.5 Heterophenomenological Reduction

The point of reconstructing subject's heterophenomenological world is to see clearly and univocally what is to be explained. For if the task is to explain the relation between brain states and mental states, we have to know both *relata*. The method of heterophenomenology is supposed to be for mental states the same as what brain-scanners and other devices are for brain states:

Heterophenomenology is the beginning of a science of consciousness, not the end. It is the organization of the data, a catalog of *what must be explained*, not itself an explanation. [Dennett, 2005, p. 40]

While there is no difficulty as to the objectivity and univocity of the “hard” scientific data (at least before they are interpreted), there is a threat of substantial equivocity in subjective reports on one's mental states. For this very reason, Dennett proposes to abandon *autophenomenology* (introspection, reflective contemplation on one's own mental states) and do *heterophenomenology* instead - that is to treat subjective reports from the third person perspective. It might be objected that *heterophenomenology* will leave something out, namely those mental states that we regard as ineffable, such as the feel of existential anxiety, for example. We might, of course, try to describe it in our idiosyncratic way, but then we run the risk of being misapprehended, for the more idiosyncrasy is involved, the less intersubjectively comprehensible it is. What does Dennett propose to do with such incommunicable experience? He would say we should first try to pay attention to other than verbal expressions of these peculiar mental states (such as the heart frequency, sweating, etc.) or just skip over them, because there is no way to grasp them²⁰. The ultimate possibility of the occurrence of a really incommunicable mental state does not render heterophenomenology useless, in Dennett's eyes. Some theorists (Nagel, Chalmers, Searle, for instance) insist that heterophenomenology cannot do justice to our experience, for it is,

²⁰This is a fine example of Dennett's consistent application of verificationist stance to mental states. Particularly Wittgenstein's “whereof one cannot speak, thereof one must be silent” could be the motto of his approach to consciousness.

by definition, third-personal, and as such, it possibly leaves something out. The reason is supposed to be the following:

If the subjective character of experience is fully comprehensible only from one point of view, than any shift to greater objectivity—that is, less attachment to a specific viewpoint—does not take us nearer to the real nature of the phenomenon: It takes us farther away from it. [Nagel, 1974, p. 447]

Dennett doubts the antecedent of this conditional²¹, and to the claim that heterophenomenology does not deal with experience *as such*, but merely with reports about it, he would reply that he is not at all sure whether there *is* something like experience *as such*. Of course he does not deny he experiences things (whatever it means), he only denies the claim that there is a privileged access to one’s experience from the first-person point of view. He defends heterophenomenology as follows:

Heterophenomenology is explicitly not a first-person methodology (as its name makes clear) but it is also not directly about “brain processes and the like”; it is a reasoned, objective extrapolation from patterns discernible in the behavior of subjects, including especially their text-producing or communicative behavior, and as such it is *about* precisely the higher-level dispositions, both cognitive and emotional, that convince us that our fellow human beings are conscious. [Dennett, 2005, p. 149]

What makes some mental states intelligible from the third person perspective? It is the *first-person plural presumption*, as Dennett puts it. We assume the others have the same “things” in their stream of consciousness as we do - such as pain, anger, joy, etc. Dennett points out that this presumption might be misleading since we are possibly not so much alike in the way we experience things as we think we are; in other words, our idiosyncrasies might be substantially incompatible²². What to rely on, then? Dennett’s choice, though never explicitly mentioned, is to rely on the informational aspect of a mental state. I think he would accept the claim that what makes two mental states alike is the information these mental states entail²³. It perfectly fits in Dennett’s functionalist view of mind and moreover,

²¹Cf. [Dennett, 2005, p. 36]

²²Cf. [Dennett, 1991, p. 67]

²³More precisely, two mental states are identical (in type, not numerically, of course) insofar as they play the same functional role in the cognitive mechanisms of both bearers of the mental states. This amounts to the same, if I understand Dennett correctly, for the information entailed in a mental state is nothing but the functional specificity of that mental state

it justifies the “reduction” of experience to speech acts that are, in principle, of propositional and hence informational character. It doesn’t follow that endorsing this position implies an eliminativist attitude to the subjective, non-informational character of mental states. Despite Dennett’s bold claims that our mental states are *nothing but* reactive dispositions, we may maintain agnosticism as to the intrinsic nature of mental states and yet pursue the project of heterophenomenology. Consider a simple case of pain: you feel a sharp pain in your shin and so do I; when the pain subsides, we may try to describe to one another how it felt. Certainly, we may never be sure that we felt the same thing, we cannot compare it in our view, in a view of a single person. However, we can both agree that the pain entailed pieces of information such as: “There is tissue damage in your left shin.” “Something has just hit you in your left shin,” etc. Intuitively, it seems that this informational aspect of pain does not at all capture the important thing that makes it being a pain, the painfulness, but it is not at all obvious that “painfulness” of pain is ensured by some intrinsic property of the corresponding mental state²⁴. I will get back to this question later; for the time being, let’s see what implications this emphasis on the informational aspect involves.

2.6 Incorrigibility of the First Person

First of all, it seems, and Dennett explicitly claims it²⁵, that the traditional idea of infallibility or incorrigibility of the first person perspective falls. In what sense, however, does it fall? Dennett supports his claim by pointing out that many people believe their visual field is uniformly detailed and focused, even though “in fact” it is not. He illustrates this by a simple experiment: a subject is asked to pick a card, hold it out at the left or right periphery of his or her visual field (while constantly looking at a certain spot so that he or she does not see the card directly)²⁶. The subject will be able to tell neither the card’s number, nor its colour. Does it mean that the subject of this experiment falsely believed that his visual field was uniformly detailed? I don’t think so, for what the subject referred to by the words “visual field” was certainly something different than what Dennett had in mind. Whereas Dennett, a man of science as he is, used a scientific term that may denote the

²⁴Since the example of pain is often used by those who are in favor of the term “quale” and insist on its relevance, Dennett, who dismisses the term and all its implications, rose to the challenge and attempted to explain “painfulness” of pain while not appealing to any intrinsic property. For his explanation, see [Dennett, 1997, p. 214-223]

²⁵Cf. [Dennett, 1991, pp. 67, 319, 359] At this point, Dennett draws on the work of Gilbert Ryle who argued that the idea of “the privileged access” is wrong.

²⁶Cf. [Dennett, 1991, p. 53-54]

2D image as represented at the retina²⁷, the subject, a layman, took the visual field to be just what he sees – he might have understood Dennett’s question (“Is your visual field uniformly detailed?”) as: “Is the representation of what you see uniformly detailed?” Is it simply a mistake in reference that occurred on the part of Dennett and the subject? We tend to think that both Dennett and the subject are right: (1) Dennett is right because he refers to the retinal image and it is indeed less focused and black-and-white at the periphery, (2) the subject is right because he or she refers to his or her representation (that was constituted in the meantime between the excitation of the retina and the assertive utterance) which is indeed such that it seems uniformly detailed.

What might dissolve this puzzle is rephrasing the original question in terms of information availability: “Do you think you dispose of the same richness of information both in the center of your visual field and at the periphery?” It is a clumsy question to ask, but it points out what we are corrigible of. We can be shown that some information is not available to us, although we thought it was (the visual field example), and vice versa – we can be shown that we dispose of information we claim not to be conscious of (the blindsight phenomenon²⁸). But we are incorrigible as to how things seem to us. Dennett concedes:

You are *not* authoritative about what is happening in you, but only about what *seems* to be happening in you, and we are giving you total, dictatorial authority over the account of how it seems to you, about *what it is like to be you*. And if you complain that some parts of how it seems to you are ineffable, we heterophenomenologists will grant that too. What better grounds could we have for believing that you are unable to describe something than that (1) you don’t describe it, (2) confess that you cannot? Of course you might be lying but we’ll give you the benefit of the doubt. [Dennett, 1991, pp. 96-97]

So the objects of heterophenomenology are subject’s reports on how

²⁷I have made up this definition, it may actually be otherwise. Nonetheless, even if the definition was in fact different, it wouldn’t affect my argument.

²⁸People that have part of their visual cortex damaged for some reason suffer from partial blindness to that area of their visual field that corresponds to the damaged part of the visual cortex. Nevertheless, if forced to guess about whether a stimulus is present in their blind field, some patients do better than chance while insisting they see nothing (they are not conscious of it). They are also able to perform some sophisticated actions, which cannot be explained unless it is accepted that the information about the stimulus is available to some parts of patient’s cognitive system. For a philosophical discussion of the blindsight phenomenon, see, for example, chpt. 12: “More about blindsight” in Humphrey, N.: *A History of the Mind*, 1992.

things seem to him. Dennett willingly gives the subject total authority over this matter, for the reports are mere fictions²⁹ – the seeming the subject experiences is not real:

You [a fictitious partner in a dialogue] seem to think there's a difference between thinking (judging, deciding, being of the heartfelt opinion that) something seems pink to you and something *really seeming* pink to you. But there is no such phenomenon as really seeming – over and above the phenomenon of judging in one way or another that something is the case. [Dennett, 1991, p. 364]

Dennett here dismisses, in a kind of mockery, the idea of *real seeming*, which refers to the assumption that whenever there is an object seeming to us, this action of seeming is a real event in the world – as if the object was *really* projected on a screen that is being watched by our mind. This metaphor opens up the way to consideration of the standard model of consciousness which Dennett calls the Cartesian Theater Model, and which he strictly rejects.

3 Replacing the Cartesian Theater

One of the first steps Dennett makes is the rejection of substance dualism. As I have mentioned above (p. 10), the main reason is that dualism leads to the mysterianist attitude to the mind-body problem in light of which the whole scientific approach is deemed to be pointless. Dennett shows that the initially reasonable idea of Descartes's, that since his thinking and his body have essences apparently independent on one another, they must be founded in different substances, turns out to be less appealing when it comes to the problem of causality between the two substances. The infamous pineal gland, the place where the body informs the soul and the soul issues commands to the body, seems to be the weak point of Descartes's theory, and it exploits a fundamental intuition which most of us harbour: there must be a place in the brain where the information from the body (sensations) becomes conscious. That is roughly the idea of the Cartesian Theater – it is a place in the brain where every conscious mental event is assembled from unconscious clusters of information sent by sensory organs or other cognitive centers in the brain. The name is derived from the metaphor of a stage and an audience: Cartesian soul observes the performance played by senses at the pineal gland and

²⁹Cf. [Dennett, 1991, p. 97]

thereby the soul gets its sensations, or better say, the mind becomes conscious of the contents staged at the Cartesian Theater. Such a metaphor, that is primarily supposed to explain cognition in the framework of cartesian (or generally dualistic) ontology, is obviously fallacious, since it explains perception (here as a process whereby the immaterial soul becomes conscious of sense-data provided by material sensory organs) in terms of another perception, namely that of the soul at the Cartesian Theater. Exposed this way, the idea of Cartesian Theater is a paradigmatic case of a homunculus argument of which Gilbert Ryle warned in his *Concept of Mind*.

Dennett's critique, however, is not focused on substance dualists that maintain the idea of the Cartesian Theater, for the idea is so obviously wrong that there can hardly be any dispute about it. Dennett's anti-cartesian argument is aimed at materialists in cognitive science, neurophysiology etc. that fail to see they themselves have been lured into the trap of the Cartesian Theater. Dennett calls them *cartesian materialists* and their distinctive feature is that they more or less explicitly presuppose a specific place in the brain where every incoming information becomes conscious. This idea is, according to Dennett, no less fallacious than the original metaphor of a stage and an audience, and his argument disclosing the disguise of the Cartesian Theater (henceforth the CT) in works of some materialists deserves attention.

3.1 Rejecting the Cartesian Theater

Naturally, we think of our experiences as happening in time and thus as possibly localizable at a point at the timescale; moreover, if we are materialists, we believe they can be localized in the brain. If I drive a car and evade a passer-by suddenly crossing the street, there must be a moment at which I realize the presence of the passer-by in the road. My eyes send the visual information to the visual centers in my brain, and when it is processed and sent further, it is realized by me. Thereupon the appropriate commands are issued to my muscles to steer to wheel and evade the passer-by. It seems that the moment of the realization must happen when the afferent signal changes to the efferent signal. But how could we find such a place of divide? As Dennett remarks, "if we could say exactly where the experience happened, we could say when it happened, and vice versa" [Dennett, 1991, p. 107]. Neurophysiologists haven't tracked down any common place in the brain that would be active in the meantime of many different stimuli and reportedly conscious actions. Of course it may have rested hidden, but the core of Dennett's argument is different: were there actually such a place, where the information becomes conscious, then the order of our experience would be the order in which the pieces of information would enter the place.

Discussion of a special case of the *phi* phenomenon ³⁰ shows that it cannot be this simple way.

To put it in a nutshell, the *phi* phenomenon is a perceptual illusion in which a perception of motion is produced by a succession of still images. In the special case discussed by Dennett, there are two spots flashing in succession, a red and a green one, which results in the illusion of a moving spot that changes its color from red to green (given the red spot flashed first) in the middle of its trajectory. Naturally, the subject perceives it as though the moving spot first changed the color and then landed at the final position (which is that of the green spot). If we stuck to the idea of the CT, we would have to hold that the information of the change of the colour arrived to the CT before the sensation of the green spot happened. But the colour to which the moving spot changes clearly depends on the colour of the second flashing spot, so unless we see the second spot, we not only don't know the next colour, but we also don't know there is anything moving.

One might insist that the visual centers delay the incoming information for a while so that they can, in special cases like this, make up a plausible, though illusionary, perception out of different sensations and subsequently send it to the CT. Dennett takes this possibility seriously, and he develops two models of such illusion-making process: *orwellian and stalinesque revisions*. He convincingly shows that at the level of brain processes there is no way to decide whether an illusion was induced by an orwellian (retroactive) or stalinesque (prospective) revision. Yet if there were the CT, there should be no problem finding out which revision has occurred³¹. This supports, according to Dennett, the view that the idea of the CT in the brain is wrong, and both models are dismissed as insufficient means to the explanation of the *phi* phenomenon.

The *phi* phenomenon indicates that we are liable to represent very short diachronic events in a made-up temporal order³². As we represent objects as

³⁰For the whole discussion and the related argument see [Dennett, 1991, pp. 120-132].

³¹The argument is rather long and philosophically irrelevant in details. For this reason I haven't incorporated it in my paper. However, it substantially contributes to overall understanding of Dennett's theory. For the detailed discussion of orwellian and stalinesque revision see [Dennett, 1991, pp. 115-126].

³²This statement entails a particularly strong epistemological belief that we dispose of means that enable us to tell reality from illusion. Although this epistemological problem should not be ignored completely, such problems are temporarily set aside for good reasons. Still I think there is a meaningful way to speak about reality in this case: the experimenter can be said to be the arbiter of what counts as reality (as far as the experiment only is concerned), since the conditions of the experiment are under his control. In our case, the experimenter designs the whole setting so that there are two spots ready to flash in succession. If there is the external world and if there is a method that can be reasonably

being localized in space, so we place events on our imaginary timeline, though the information about them might have entered our brain in different order. In Dennett's words: time of representing is not the time represented³³. This distinction is useful only at the temporal microlevel of brain processes, i.e. tens or hundreds of milliseconds; events of greater span than a second are represented unproblematically. The point of the experiment is not to support the theory of the Grand Illusion which suggests that we never perceive the world as it *really is*, but to show that it is hopeless to search for the bottleneck in the brain that would explain the apparent linearity of our conscious experience as arising on brain processes that run in parallel. Such a bottleneck, through which all the pieces of information would have to flow one by one, would not only be responsible for the represented order of events, but it would also become the most researched area in the brain, the material CT.

It is worth noting that Dennett does not survey nature of the time of consciousness, at least not in the way Husserl or Bergson did. He doesn't regard temporality as an intrinsic and inseparable feature of consciousness as Phenomenology tends to. For him, temporality of consciousness is manifested solely in the tendency of representing events linearly ordered. It is not at all surprising if we consider his functionalist attitude. In fact, Dennett would even refrain from talking about the time of consciousness: there is no time in consciousness, no temporal flux of thoughts, there is only a representation of temporal events and their relation to each other (before – after), which ultimately makes the impression of the time. In Dennett's view, time in consciousness is merely a data item of most (perhaps not even all) of our mental contents – it is a piece of information. Thus *time in consciousness* is only a representation of things (events) as ordered by relation “before – after”³⁴. On the other hand, there is *consciousness in time* which means that

called objective, then we may rightfully claim that the experimenter has the privileged access to the reality of the experimental situation because it largely depends on his setting.

³³It is perhaps better to rephrase it as “order of representing is not the order represented” because the time of consciousness is definitely not the same in nature as the physical time. If we want to compare these two, the only criterion we can use as a measure is the relative order of both (provided they are both as linear as we hold them to be). An isomorphism between the two timelines would then count as a sign of true representation of the happenings in the external world.

³⁴Surely, this representation of temporal order must be physically realized, for example in some neural circuit, and hence finds itself in “real” time, but the real time when a bunch of neurons began representing a point at a fictitious timeline is irrelevant to the time represented by the neurons. If some neurons rewire in my brain right now so that they become representing a moment in the past, I may, for example, suddenly believe that I finished this paper a year ago – and I would equally believe that the information

consciousness as a biological process (or, to put it in Dennett's terms, as a running algorithm that is being realized by a biochemical Turing machine – the brain) takes place in the physical time.

We have grown sophisticated enough to recognize that the products of visual perception are not, literally, pictures in the head even though *what they represent* is what pictures represent well: the layout in space of various visible properties. We should make the same distinction for time: *when* in the brain an experience happens must be distinguished from when it seems to happen. . . . The representation of space in the brain does not always use space-in-the-brain to represent space, and the representation of time in the brain does not always use time-in-the-brain. [Dennett, 1991, p. 131]

This interpretation may seem to be stretched too far, but the purpose is to draw attention to the important role of representation in Dennett's model. I hope I have shown that Dennett strictly discriminates the time of representing and the time represented, where the former is the physical time accessible by objective means, while the latter is the index attached to most of our mental events and thus accessible within consciousness³⁵.

3.2 Introduction of the Multiple Drafts Model

Since it is out of dispute that the brain processes information in parallel, and since there is no CT to be found in the brain, the problem becomes how parallel and spatially scattered brain processes make up the seemingly unified consciousness. Indeed, the temporal parallelism and spatial extension of brain processes seems to be incommensurable with the linear and spaceless character of consciousness³⁶. This incommensurability resembles that of a sign and its meaning. Every sign (as a token) is materialized, yet its sense seems to be spaceless and non-material. These two cases of incommensurability are in fact two sides of the same coin, for there is no meaning without

that I finished the paper is a year old too, though it has been physically realized only few moments.

³⁵Such a crude elaboration of the kinds of time seems to be hopelessly fallacious, for one could reasonably object that we “meet” the time only in consciousness and that the data from the so-called objective means of measuring time have nonetheless to pass through consciousness where they necessarily lose their objectivity (in the sense of viewer-independency). I don't intend to go on with contemplation on time in Dennett's terms, because I know well that the more one thinks over time, the more complicated it gets and it is not, in the end, so important for this paper.

³⁶This point is often stressed by anti-reductionists, Searle and McGinn, for example.

a meaner, no word without a reader, no sense without an understander (at least from the nominalist point of view) and these are all conscious beings. Meaning of a sign cannot be detached from the intentional structure of the act of understanding at the side of the sign-user. The purpose of a sign is to refer to something else, and the reference is always in the eye of the beholder. It is an object's appearing to us as representing something else that makes it a sign; a pure sign that would represent something *by itself* is an abstractum. Hence, a sign is spaceless and non-material only as an abstractum, i.e. when it is understood. And understanding is perhaps the central feature of consciousness. The Cartesian Theater served as the place of representation and there were all three constitutives of the representational act: I as the audience³⁷, the flow of information as the performance, the bits of information as the content of the play. Therefore if Dennett rejects the CT, he has to show how understanding, representation, and its intentional structure³⁸ occur upon parallel brain processes. As will be seen later, Dennett's stratagem is to explain the self as an illusion and the whole intentional structure along with it - he tacitly assumes that the intentional structure loses its relevance as soon as the concept of the self is dissolved.

Even if Dennett proves the self is an illusion, he has to show why it seems (to us) there is something like the self. But then again, there is the *seeming* to be explained. As Descartes first observed, the object of seeming may be illusory, but the act of understanding the meaning of what seems to me is indubitable.

3.3 The Multiple Drafts Model

The Multiple Drafts Model (henceforth the MDM) is supposed to replace the old CT model of consciousness. Dennett himself points out several times³⁹ that the MDM is so counterintuitive that it is very hard to grasp at the first reading, especially if we are used to think about consciousness in dualistic framework. He intends to clarify the initially incomprehensible theoretical model by discussing special cases of perception, like the *phi* phenomenon,

³⁷Introducing an 'I' or a self in the CT model leads directly to the infamous Homunculus Fallacy: "How does the self as the audience in the CT understand the performance?", "How does the self know what the bits of information mean?" If not for other reasons, this compels us to reject strictly the CT model. However, we then face the challenge of explaining the faculty of representation in different terms.

³⁸By the intentional structure of representation I mean the following: every representation is conceived by (1) *a subject* (a self) (2) through an *act of understanding* whereby the subject intentionally relates himself to (3) the *represented object*. Hence: (1) *I* (2) *represent* (3) *something*.

³⁹Cf. [Dennett, 1991, pp. 113, 227, 321]

whereby he shows that the MDM explains those cases satisfactorily, unlike the old CT model. As to me, the MDM is not so obscure because it is intellectually difficult to grasp, as Dennett seems to claim, but rather because it is only a model of how consciousness works; it expounds the structure of consciousness, but it fails to explain *in causal terms* why a subject is conscious of this and that. This is not necessarily a drawback, especially if consciousness as we know it is excluded from the causal chain of physics, which seems to be a result of Dennett's theory of consciousness. Nonetheless, we intuitively judge a full-fledged causal explanation of a phenomenon as more convincing than a dissolution of the phenomenon as a mere illusion. Recalling the subsection "Nature of the Explanation" (p. 12), the MDM cannot answer the question of the kind "Why is that so?" and as an explanation it would be classified to the "scientific" kind of explanation. On the other hand, the CT model offers a causal "pseudoexplanation" that just avails of a mysterious kind of mind-body interaction (the observation at the CT) which, at the end of the day, renders it unsatisfactory. Despite of the mysterious causation involved, the CT model is closer to the "empirical" explanation than to the "scientific," and that is one of the reasons, I think, why it is so hard for Dennett to fight against it with his own model – for "scientific" explanations usually make non-scientists, philosophers notwithstanding, feel deprived of some substantial answers (What is the gravity?), whereas the "empirical" explanation, if successful, seems always to be complete, since we are usually familiar with all the principles and things involved (we know by experience that objects fall if unsupported, but we know planets rotate around stars only because we are taught so, because it follows from physical principles that are nonetheless abstracted from our everyday experience).

Since the MDM is exposed via series of discussions of special cases of perception, it is hard to make a brief and concise exposition of the MDM without being susceptible of omitting important features. There is, however, a theoretical skeleton to be found which is only fleshed out by concrete examples. The best point to start at is the outline of structure of information processing in the brain:

According to the Multiple Drafts model, all varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous "editorial revision." [Dennett, 1991, p. 111]

This is still nothing new, and even many cartesian materialists would agree, provided this brief outline does not exclude the possibility of the ul-

timate place in the brain where information becomes conscious (the CT). In fact, the claims in this outline follow naturally from consideration of the architecture of the brain. It is indeed such that the brain continuously processes information from many sources and does so at many different centers. Therefore, unlike with digital computers that process information serially, the parallel processing makes it virtually impossible to tell where (and consequently also “when”) this or that mental content takes place. Suppose you see a snake. A complex, yet unified, phenomenon arises in your consciousness. A part of this phenomenon is the image of the snake, another part is the knowledge that it might be poisonous; you also might be aware, while watching the snake in stupor, of your alertness manifested in overtensed muscles ready to run. All these pieces of information are present in the unified phenomenon of snake-seeing, yet there is no place in the brain where all this has come together, there is no CT staging the snake-seeing play. The phenomenon “physically” (i.e. from the materialist’s point of view) happens at several places in the brain at once: the image is somehow represented in the visual cortex and associated centers, the fear originates, say, in amygdala, etc. However, we should not go as far as to assume that though the overall information is scattered, its pieces can be precisely localized and literally said to be here or there⁴⁰. The reason is that, thanks to the physical structure of a neuron and thereon dependent logical structure, every neuron works primarily as a discriminator of certain features. Information then is represented subsymbolically in a dynamic pattern of activation of a group of neurons the work of which consists in finer and finer discrimination.

Discrimination is the main process of the MDM:

What we actually experience is a product of many processes of interpretation – editorial processes, in effect. They take in relatively raw and one-sided representations, and yield collated, revised, enhanced representation, and they take place in the streams

⁴⁰And even if we could, we would have to find an answer as to how it comes that these pieces of information are available to “the self” at once. Dennett will say that the self is a fictitious character “reconstructed” upon the flow of narratives of information – our selves come to being by much the same process as when we read somebody’s autobiography (we are constantly exposed to the flow of information concerning our body and its interests – this narrative gets unified by the useful, yet fictitious, character of the self). Dennett posits the self to be *the center of narrative gravity*. His answer to this point may seem to be rather cunning at the first sight, but it really follows from his overall theory. Information about his account of the self as the center of narrative gravity can be found in his article “The Self as a Center of Narrative Gravity” in F. Kessel, P. Cole and D. Johnson, eds, *Self and Consciousness: Multiple Perspectives*, Hillsdale, NJ: Erlbaum. Also: [Dennett, 1991, pp. 413-430].

of activity occurring in various parts of the brain. This much is recognized by virtually all theories of perception, but now we are poised for the novel feature of the Multiple Drafts model: Feature detections or discriminations *only have to be made once*. That is, once a particular “observation” of some feature has been made, by a specialized, localized portion of the brain, the information content thus fixed does not have to be sent somewhere else to be rediscriminated by some “master” discriminator. In other words, discrimination does not lead to a representation of the already discriminated feature for the benefit of the audience in the Cartesian Theater – for there is no Cartesian Theater. [Dennett, 1991, pp. 112-113]

This passage finally opens up the counterintuitive part of Dennett’s theory. The main point stems from the conviction that since there is no CT, it is hopeless to look for a specific place or a process that makes some information conscious. Dennett then concludes that once a feature is discriminated by neurons, it becomes available to consciousness as a fixed mental content. However, to be available does not mean to be a part of, and therefore one may ask: “What makes a discriminated feature be in the stream of consciousness?” Dennett regards it as a misleading question:

It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience, and it is a confusion, as we shall see, to ask *when it becomes conscious*. These distributed content-discriminators yield, over the course of time, something *rather like* a narrative stream or sequence, which can be thought of as subject to continual editing by many processes distributed around in the brain, and continuing indefinitely into the future. This stream of contents is only rather like a narrative because of its multiplicity; at any point in time there are multiple “drafts” of narrative fragments at various stages of editing in various places in the brain.

...

Most important, the Multiple Drafts model avoids the tempting mistake of supposing that must be a single narrative (the “final” or “published” draft, you might say) that is canonical – that is the *actual* stream of consciousness of the subject, whether or not the experimenter (or even the subject) can gain access to it. [Dennett, 1991, p. 113]

Dennett here introduces the notion of a narrative which is to be understood as a flow of information that nevertheless entails no subjectivity, insofar it is possible for information to be independent of any subject that understands it⁴¹.

3.4 Continuity of (the notion of) Consciousness

As to the structure of consciousness that the MDM implies, the main point is that *at some level* there is plurality of information streams that continually change their content, originate, and fade away again⁴². The level at which “consciousness” has the proposed structure is nevertheless very disputable, since many theorists (all Phenomenologists, I assume, and Searle, Chalmers and others) would object that a distinctive feature of consciousness is its linearity and unity, and therefore they would hold that the asserted plurality belongs to some pre-conscious level. This disagreement, I think, does not originate from radically different views on the structure of consciousness, but rather from different usage of the notion of consciousness, which, in the end, is caused by different attitudes to the mental. Whereas Dennett’s operationalism and disguised behaviourism makes him treat mental states in terms of the overt acts they lead to, all the above mentioned theorists tend to start their considerations of consciousness at mental states which are supposed to be given as facts, together with their semantic and subjective character. Consequently, consciousness is for them a domain of either-or: either you are conscious of something at a moment or not; and the only authority is you and your ability of reflexion. For Dennett, on the other hand, consciousness is a domain with continuum of degree the value of which depends on the intensity of the overt act⁴³. Dennett himself remarks:

⁴¹From Dennett’s way of talking about the narrative I gather he would deny that his notion involves any subjectivity. He would surely deny the presence of intrinsic semantics, which is what Searle appeals to very often in his works. In my opinion, the only sense in which “meaning” might be ascribed to the narrative consists in the potential to cause an overt act of the subject. For instance, the meaning or, say, a semantic character of a feature discrimination done at amygdala that is responsible for the feel of fear in the snake example consists in its potential to induce a kind of emergency status thanks to which the body is ready to act quickly and correspondingly.

⁴²As far as I can understand the proposed structure, once we open eyes, for example, the perception of visual stimuli gives rise to a “visual” narrative that fades away if we close the eyes again and which can flow parallelly to the “auditory” narrative and many others. This, however, is not as important as the reason why we all intuitively regard the stream of consciousness as unified and linear.

⁴³In this interpretation then, I am more conscious of pain that makes me scream than of pain that makes me only rub the hurting limb. On the one hand it seems nonsensical to quantify consciousness of mental states, on the other hand we would agree that great

The absolutist or essentialist philosopher is attracted to sharp lines, thresholds, “essences” and “criteria.” For the absolutist, there must indeed have been a first mammal, a first living thing, a first moment of consciousness, a first moral agent;

...

Opposed to this way of thinking is the sort of anti-essentialism that is comfortable with penumbral cases and the lack of strict dividing lines. Since selves and minds and even consciousness itself are biological products (...), we should expect that the transitions between them and the phenomena that are not them should be gradual, contentious, gerrymandered. [Dennett, 1991, p. 421]

In short, Dennett tells us that consciousness is not as we know it to be: “I don’t maintain, of course, that human consciousness doesn’t exist; I maintain that it is not what people often think it is” [Dennett, 2005, p. 71]. What is the sense, however, of saying that consciousness is not what we think it is? Is not consciousness after all precisely that what we experience, what we are most intimately familiar with? Not so for Dennett, who focuses on what consciousness does, rather than how it seems. Recall again the section on heterophenomenology. We can imagine that Dennett first gathers all the data (both the speech acts and the brain-scans) and then ponders: Well, the subject reports to be unconscious of this and that information, the brain-scans, on the other hand, indicate that the information is likely to be there in the brain; so what must be consciousness like, if this is what it does? Since there is no privileged access to consciousness (as it *really* is; everyone has privileged access only to how it *seems* to him), we all dispose, in principle, of the same information from which we can deduce what consciousness is like. Maintaining this view, Dennett can go as far as to say that consciousness is not unified and continuous, rather the opposite⁴⁴.

Despite the plurality of narratives, we still experience a single stream of consciousness which moreover exhibits potential linearity, that is to say we feel we could in principle always determine whether a given mental content precedes or succeeds another mental content. The perception of the *phi* phenomenon is experientially linear, though the experienced succession of

pain pulls other things out of our phenomenal space though we might still rightly be said to be conscious of them. A reasonable way out of this difficulty is to make a distinction between consciousness and awareness so that they would capture both senses in which we say we are “conscious” of something, but Dennett uses the term “awareness” very rarely, let alone systematically.

⁴⁴Cf. [Dennett, 1991, p. 356]

images of flashing dots originated from parallel information processing. The task for Dennett is then to explain why the stream of consciousness *seems* unified and linear to the subject. He cannot explain it causally in terms of a special process whereby the plurality of narratives is united, for that would be a regress to the CT model. His strategy is, I think, twofold: primarily, he explains it away as an illusion originating in our mistaken concepts of mental states, in much the same way as Gilbert Ryle put in doubt the usual way of thinking of mental states by drawing attention to category mistakes present in many preceding theories of mind. Secondly, he argues from the evolutionary stance that the complex “illusion” of the self is advantageous, and he narrates “an evolutionary myth” about the origin of consciousness based on the development of language.

3.5 Making Contents Conscious

To help us understand how and why the MDM resists the belief in a clearly delineated extension of consciousness (i.e. the set of all conscious mental states), Dennett expounds the analogy of the process of publishing which actually gave name to the MDM. Whereas earlier “it used to be that virtually all of an article’s important effects happened *after* appearance in a journal and *because of* its making such an appearance” [Dennett, 1991, p. 125], nowadays is the situation different:

With the advent of word-processing and desktop publishing and electronic mail, it now often happens that several different drafts of an article are simultaneously in circulation, with the author readily making revisions in response to comments received by electronic mail. Fixing a moment of publication, and thus calling one of the drafts of an article the *canonical* text – . . . – becomes a somewhat arbitrary matter. Often most of the intended readers, the readers whose reading of the text matters, read only an early draft; the “published” version is archival and inert. [Dennett, 1991, p. 125]

The first part of the analogy concerns the structure and editorial processes that have already been discussed. The second part, mentioning the arbitrariness and the impact of an article before its publication, is explicated as follows:

Similarly – . . . – if one wants to settle on some moment of processing in the brain as the moment of consciousness, this has to be

arbitrary. One can always “draw a line” in the stream of processing in the brain, but there are no functional differences that could motivate declaring all prior stages and revisions to be unconscious or preconscious adjustments, and all subsequent emendations to the content (as revealed by recollection) to be post-experiential memory contamination. [Dennett, 1991, p. 126]

As it is arbitrary to say that the final published version is the main text, for the most important effects had been caused before it was published, and moreover by many different drafts, so it is arbitrary to say when a discriminated feature becomes conscious. Since the effort to look for a moment of transition from unconscious to conscious goes in vain, it is more useful to look for the effects of a feature discrimination. If a growing patch in the visual field is discriminated as an object looming at one’s face, one of the possible effects might be taking evasive action. Sometimes the reaction to a stimulus is so swift that we realize the presence of the stimulus only after we have already reacted, which phenomenologically means we became fully conscious of it as mere passive spectators, while the appropriate measures had been taken earlier “pre-consciously”⁴⁵. Interestingly, we could not tell whether we lately became conscious of the stone because “it was just there, in our visual field,” or because we were surprised by our very reaction. It could possibly be that while strolling around you suddenly found yourself swiftly moving head to the left with no previous deliberation involved. If your body surprises yourself by such an unexpected movement, it is a good enough reason to look for the cause – and this search for the cause may contribute substantially to the (conscious) realization that a stone has been looming at you. This example is only to illustrate the importance of all kinds of effects and their possible contribution to awareness.

However, this would not convince the theorists who tend to regard consciousness as the domain of either-or. If mental contents are rightly said to be⁴⁶ either conscious or unconscious, then there has to be a sharp boundary between the two states. Let’s now ask what this boundary could consist of,

⁴⁵This is enabled by the fact that there are several pathways between the retina and the rest of the brain. The evolutionary older pathways can transmit the signal directly to the centers responsible for instinctive self-preserving mechanism such as dodging, which can process the incoming information faster than the visual cortex.

⁴⁶Perhaps I should write “if mental contents *really are*,” since according to those theorists the consciousness of a mental state is a matter of fact, and not only a matter of right usage of mental conduct terms (Ryle). Dennett often stresses that those who believe in matter-of-factness of consciousness want to preserve the reality – appearance distinction for consciousness, the distinction between *real* seeming and apparent seeming. See p. 24 or [Dennett, 1991, p. 131].

phenomenologically rather than physically. As I sit here at the desk, I am not conscious, in the either-or sense of it, of the wall behind me, nor am I conscious of the painting right in front of me, since I do not relate myself to it intentionally (although the sense-data of it are surely available in the brain, and I would readily become conscious of it if, for example, the painting suddenly changed). On the other hand, I am conscious of what I am writing, and of the computer I am using as well. The main difference between the conscious and the unconscious, I assume, is twofold: (1) the unconscious contents are never articulated “in my head,” whereas the conscious often are (though not always); (2) the conscious contents are always objects of intentional relations⁴⁷. The second point concerning intentionality is often stressed by antireductionists like Searle, Nagel, et al., who often regard it as an intrinsic and distinctive feature of consciousness whereat the explanation of how consciousness works should start. Intentionality indeed cannot be dismissed as irrelevant, yet it is an open question whether it is an intrinsic property of the brain as a physical machine with causal powers (Searle), or a side-effect arising on symbol manipulations.

Dennett seems to focus on the first point. The articulation of a content in language and more importantly its presence in memory is what makes it “conscious.” He claims that

what happened (in consciousness) is simply whatever you remember to have happened. The Multiple Drafts model makes “writing it down” in memory criterial for consciousness; that is *what it is* for the “given” to be “taken” – to be taken one way rather than another. There is no reality of conscious experience independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory). [Dennett, 1991, p. 132]

The last sentence represents the core of Dennett’s operationalist attitude to mental states, and it is also the main reason why he rejects the philosophical notion of *quale*. But does not Dennett here try to take our attention away from the main issue – namely how the phenomenal single stream of consciousness arises upon multiple narratives? The claim that “what happened in consciousness is whatever you remember to have happened” may be considered as trivial, since every conscious content leaves a memory trace, no matter how long it lasts. It is questionable what kind of memory does Dennett have in mind, but it surely is rather the short-term memory (or “working

⁴⁷Who or what relates intentionally to the conscious contents is a hard problem of its own. I avoid mentioning the subject of the intentional relation because I don’t want the phrase to entail ontological commitment to the self, or transcendental ego, which has played this role in the philosophical tradition.

memory” if you want) than the long-term memory, because he supports his claim by discussing examples where a subject reports on events immediately after they have happened. The short-term memory then could be the place where contents are explicitly conscious, but since Dennett does not explain how the memory (or the global workspace, as he sometimes calls it⁴⁸) works, it seems only to postpone the explanation from consciousness to memory and thus to be no more illuminating.

Significantly, Dennett concerns himself with memory only for a short time, and he soon puts emphasis on verbal articulation of a content. Let us consider for the last time another of Dennett’s summaries of the MDM:

While some of the contents in these drafts [i.e. drafts as fragments of the narratives] will make their brief contributions and fade without further effect - and some will make no contribution at all - others will persist to play a variety of roles in the further modulation of internal state and behavior and a few will even persist to the point of making their presence known through press releases issued in the form of verbal behavior. [Dennett, 1991, p. 135]

“Verbal behavior” probably refers only to an overt utterance here, but let’s assume that an internalized speech holds the same function. Dennett’s point here is, I think, that, according to the principles of heterophenomenology, the only way to get to know there *are* some mental contents is to declare it verbally (or behaviourally, in the case of simple beliefs). So much is widely acknowledged by many theorists as long as we are interested in the recognition of *other people’s* mental contents, whereas our mental contents and, indeed, the whole consciousness, is generally considered to be transparent. What makes the ability of verbally expressing one’s mental contents important for the question of how consciousness arises is Dennett’s view which can be summarized as following: to be conscious of a mental state is to be verbally (or in other meaningful way) expressible by the mental state bearer; or, as Dennett himself states: “If I couldn’t talk to myself, I’d have no way of knowing what I was thinking” [Dennett, 1991, p. 316]. But does knowing what one is thinking amount to the same as being conscious of it? I am not quite sure, but Dennett probably thinks so, but for reasons I understand only with severe difficulties (I will nevertheless try to summarize it). His conviction is based on the idea that becoming conscious of a mental state is done by an occurrence of a higher-order mental state the content of which is the first mental state. Elevating a mental state to the higher-order mental state

⁴⁸Cf. [Dennett, 1991, p. 270]

is done by articulating the former content (this articulation is again only a kind of information processing, it cannot be conscious, for that would lead to circularity in the argument). It is as if we were becoming conscious of things by our brains talking to themselves: “The only way for a human brain to get itself into something like a higher-order belief state, we surmised, is to engage in the process rather like reporting first-order states to itself” [Dennett, 1991, p. 316]. The contribution of language to consciousness is clearly exposed in the following paragraph from Dennett’s *Kinds of Minds*:

Mental contents become conscious not by entering special chamber in the brain, not by being transduced into some privileged and mysterious medium, but by winning the competitions against other mental contents for domination in the control of behavior, and hence for achieving long-lasting effects – or as we misleadingly say, “entering into memory.” And since we are talkers, and since talking to ourselves is one of the most influential activities, one of the most effective ways for a mental content to become influential is for it to get into position to drive the language-using parts of the controls. [Dennett, 1997, pp. 205-206]

In other words, articulation of a mental content is itself a good enough behavioural effect in order for the mental content to become conscious, no matter whether the articulation is overt or just internalized, as in the case of soliloquy. To fully appreciate the meaning of this claim, however, we have to consider the gradual process of evolution which starts at unconscious and purposeless molecules and goes all the way to conscious human beings. Seeing upon which selective pressures language has evolved will help us better realize how language can contribute to consciousness.

4 The Evolutionary Myth

Basically, there are at least two lines of argumentation to be distinguished in Dennett’s work. The first is a top-down kind of argumentation where Dennett starts at consciousness as the explanandum (the folk notion of consciousness) and explains it by more basic processes and terms. The MDM is the prime example – there Dennett first scrutinizes the folk notion of consciousness and once the problems are settled, he analyzes them to simpler parts and finally he offers the complex MDM. The second line avails of bottom-up argumentation which starts at a simple and generally acknowledged fact and

by adding more and more simple features, it grows both in scale and complexity so that it, in the end, reaches the level of the explanandum (i.e. consciousness). The argument from evolution, which will be subsequently paraphrased, is of the bottom-up kind, for it starts at the level of primitive organic macromolecules and ends at the level of human being endowed with consciousness⁴⁹. The reason why I name the argument “The Evolutionary Myth” is not that I consider it an unscientific fantasy, but that it is designed mainly as a story of the origination of consciousness that helps us understand the present situation, which is after all what myths do.

The story starts at the time when “there was no teleology at all”⁵⁰ and suddenly there emerged first organic macromolecules, able to replicate themselves and feed on other organic compounds. This was the time, Dennett says⁵¹, when interests were born, for even the simple replicators can be assigned interests in self-replication, if we adopt the intentional stance⁵².

Wherever there is an interest, there is a criterion for classifying things as “good,” “bad,” or “neutral” means to serve the interest. More important, along with the interest of self-preservation there appears a point of view, i.e. a perspective from which things are judged as favorable, unfavorable or neutral, and a boundary delineating the environment from the organism, which is supposed to be the primordial origin of selfhood⁵³.

Next step is the evolution of a nervous system. At a certain degree of complexity, an organism, in order to cope, “must either armor itself (like a tree or a clam) and ‘hope for the best’, or else develop methods of getting out of harm’s way and into the better neighborhoods in its vicinity” [Dennett, 1991, p. 177]. The latter choice demands the control of one’s

⁴⁹The utility of the bottom-up argumentation is well illustrated by the law of uphill analysis and downhill synthesis as quoted by Dennett from V. Braitenberg’s *Vehicles: Essays in Synthetic Psychology* (1984): “It is much easier to imagine the behaviour (...) of a device you synthesize “from the inside out” one might say, than to try to analyze the external behavior of a “black box” and figure out what must be going inside,” quoted from [Dennett, 1991, p. 171].

⁵⁰Cf. [Dennett, 1991, p. 173]

⁵¹“The day that the universe contained entities that the universe contained entities that could take some rudimentary steps toward defending their own interests was the day that interest were born.” [Dennett, 1984, p.22]

⁵²Since all the ascriptions of interests, points of view, and reasons to obviously unconscious organisms is done from the intentional stance, we do not need to be worried by the fact that these organisms cannot appreciate their interests etc. The meaning of the claim that an unconscious organism seeks the good and avoids the bad is much the same as “it is hard-wired in that organism that certain feature discriminations lead to either avoidance or absorption.” “The interests” are built in the organism in the same way as fuses are built in a house to prevent short circuit.

⁵³Cf. [Dennett, 1991, p. 174]

activities in time and space, and

the key to control is the ability to *track* or even *anticipate* the important features of the environment, so all brains are, in essence, *anticipation machines*. [Dennett, 1991, p. 177]

To put it crudely, the further future a brain can anticipate, the better the brain is. A simple reflex of ducking a looming brick is an example of short-range anticipation, whereas the ability to abstract regularities and assign them the law-status is a higher-order faculty the purpose of which is nonetheless anticipation.

The ancestor of awareness (or consciousness, if you want) is alertness. Brains of most animals perform constant activity “on the background” that tests the presence of alarming stimuli in the environment (a pair of eyes gazing at you, for example). Once the alarming feature is discriminated, it triggers a series of reactions that finally gets the brain into an emergency status during which it searches for more information so that it performs the right action (running away, if the gaze belongs to a predator, for example). This enhanced search for information proved to be so useful that animals began to go into that mode more and more often, which soon turned into regular exploration. Information began to be acquired for its own sake, just in case it might prove valuable later. Most mammals adopted this strategy which ultimately gave rise to *epistemic hunger*⁵⁴.

Highly important is the evolution of neural plasticity, i.e. the ability to re-wire the present setting of the brain according to needs. This results in the ability to learn actively during animal’s lifetime which makes the animal and its whole kind independent to great degree of the luck of the trial-and-error method of Mother Nature’s mutations⁵⁵. Another of the important effects of neural plasticity is the faculty of representation that keeps track of a stimulus even if senses no longer attend to it. A predator able to anticipate the trajectory of prey fleeing behind high grass will be more successful than that who has to keep its eyes fixed on a prey while chasing it. The power to represent an earlier perceived object is again considered to be the foundation of the lately developed higher-order representation of abstract notions.

The last but one step on the way to consciousness is the development of language and “the habit of autostimulation.” The Evolutionary Myth has it,

⁵⁴Cf. [Dennett, 1991, p. 181]

⁵⁵Dennett devotes many pages to this step and explains it in great detail, but the main point, important for our discussion, is to illustrate how revolutionary the change from animals with fixed “hardware” to animals with dynamic “hardware” is. For the detailed discussion see [Dennett, 1991, pp. 182-193] or any article on *Baldwin Effect*.

not surprisingly, that some species developed language or less sophisticated means of communication in order to be able to share information with other animals of the same species. The primary reason for using language is supposed to have been to ask, either for help or for information. Thus every animal belonging to a language community must be ready to play the role both of the asker and the answerer in order for the communicative habits to become established. Once the communicative habits were established within a community, its members got used to ask for information whenever they needed it. This process of asking and answering then became internalized thanks to accidental autostimulation that proved to be useful:

The one fine day (in this rational reconstruction), one of these hominids "mistakenly" asked for help when there was no helpful audience within earshot - except itself! When it heard its own request, the stimulation provoked just the sort of other-helping utterance production that the request from another would have caused. And to creature's delight, it found that it had just provoked itself into answering its own question. [Dennett, 1991, p. 195]

The utility of this kind of autostimulation is based on the fact that due to less than optimal wiring of a brain, information present in one subsystem may be unavailable to another subsystem that currently needs it⁵⁶. Provided that both subsystems can access the environment, the absence of an information link between the subsystems can be overcome by sending the information to the environment so that the subsystem in need can pick it up therefrom⁵⁷. This virtual wire between the subsystems that "goes through" the environment can relatively easily become internalized as a real wire between the auditory system and the language-production center, whereby the whole process becomes private and perhaps more effective as well.

Although the origin of consciousness is not fully illustrated by the story of the development of internalized speech, it nevertheless shows how an important, if not crucial, feature of consciousness might have originated from rather meaningless, unconscious processes. Surprisingly though, Dennett nowhere discusses in detail the nature and origin of the faculty to represent things or events by words; which is what everyone would expect, considering the familiarity of Searle's Chinese Room argument among the people interested in

⁵⁶The situation when a part of the brain cannot share information with another part of the brain is not as unusual as it may seem, though it becomes clear mainly after a brain injury. The blindsight phenomenon and the cases of split-brain patients are good examples.

⁵⁷Cf. [Dennett, 1991, p. 195-196]

the mind-body problem⁵⁸. Perhaps Dennett deemed the nature of representation (and thereupon dependent semantics) unproblematic, but in the light of Searle's argument, it seems to be a desideratum for any theory of consciousness to explain how language becomes the domain of articulated sense, and consequently, how consciousness as a natural phenomenon is endowed with semantics⁵⁹. I assume that Dennett tacitly, yet intentionally, avoids this issue, partly because he considers it to be a wrongly formulated problem, for it draws on the CT model of consciousness with the infamous institution of "the central meaner," partly because he may regard meaning to be only a matter of functional relations between concepts⁶⁰ that are, in fact, more or less complex representations of things in the world as they have come in through our senses and have been subsequently altered and intertwined by our idiosyncrasies. But then again, in what sense does some neural activity represent⁶¹ this or that thing? Is it enough to say that the neural activity is a reaction to this or that stimulus and hence represents the stimulus in virtue of being causally related to it? I hope to address these questions in the section on semantics in the brain.

Although his reasons for omitting the issue of semantics may be right, it is still not enough, I think, to say that certain brain activation represents a thing in the world simply in virtue of being the reaction to the stimulus (i.e. the perception of the thing), for all the claims of the kind "A represents B" are observer-relative, in the sense that A represents B as long as there is someone who recognize A as standing for B. Or am I wrong in that? Possibly, since there might be a regress to the CT model lurking behind my argument: if every meaning must be meant by someone, then there must be something that represents *per se*, the meaner, whose natural power is to mean and who is thus the prime candidate for being the audience in the CT. Whether or not

⁵⁸Although the explicit reformulation of the Chinese Room argument in terms of syntax – semantics distinction first appeared in 1990 in Searle's paper "Is the Brains Mind a Computer Program?" from *Scientific American*, and thus Dennett might have not taken it into account when writing *Consciousness Explained*, he must have been familiar with the original formulation from "Minds, Brains, and Programs" in *The Behavioral and Brain Sciences*, (1980) where Searle already stressed the difference between handling with uninterpreted symbols and understanding them.

⁵⁹Or in the reversed order, if you hold that there is first semantics (intrinsic or not) and then language.

⁶⁰Hence Dennett possibly takes Wittgenstein's "For a large class of cases - though not for all - in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language." literally, that is as revealing the essence of meaning itself.

⁶¹As I have deduced from Dennett's works, he seems to mean by "representation" the same as what is usually called "reference" in linguistics, which is the translation of Frege's "Bedeutung."

there can be representation without a meaner seems to be the crucial point at which the ways of many theorists split, for if there must be a meaner, then semantics must be an intrinsic feature of some organization of matter (as long as we are not substance dualists), and the power to mean is a causal power of such organized matter. Thus the main disagreement between Searle and Dennett, from which all other quarrels originate, seems to be about the nature of semantics. Searle holds it to be an intrinsic property of brains as physical machines, whereas Dennett considers it to be a side-effect of the brain processes the essence of which is entirely captured in functional terms. It is tempting to juxtapose Dennett's position in the last sentence as follows: "... whereas Dennett considers it to be a side-effect of the purely syntactical program being executed by a special Turing machine - the brain," but though this view is often ascribed to Dennett, I regard it as an overstatement of what he actually says. In order to explain this, however, I have to proceed to the final part of Dennett's positive account of consciousness.

5 Memes and Virtual Machines

5.1 Memetics

The evolutionary argument doesn't end at the evolution of internalized speech, it goes further on the cultural evolution and considers its relation to consciousness. I will discuss it separately from the evolution of language and internalized speech because the cultural evolution (or evolution of memes, as we will see) takes place at different domain than the darwinian evolution, and moreover, it heavily draws on a theory that is not, as far as I know, generally acknowledged(at least not as the classic theory of evolution is).

The above mentioned theory is Richard Dawkins's *memetics*. In his paper *The Selfish Gene*, he presented an idea of a *meme* - "a unit of cultural transmission, or a unit of *imitation*" [Dawkins, 1976, p. 143]. Memes are in fact ideas or other unified and stable mental contents; the concept of meme differs from ideas or mental contents mainly in that memes are ascribed certain independency and their functioning is considered to be the same as of viruses:

Examples of memes are tunes, ideas, catch-phrases, clothes, fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperm or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation. If a scientist hears,

or reads about, a good idea, he passes it on to his colleagues and students. He mentions it in his articles and his lectures. If the idea catches on, it can be said to propagate itself, spreading from brain to brain. [Dawkins, 1976, p. 143]

What makes it reasonable to consider seriously the idea of self-propagating, on minds parasitizing memes, is that they satisfy the following conditions of evolution:

- (1) variation: a continuing abundance of different elements
- (2) heredity or replication: the elements have the capacity to create copies or replicas of themselves
- (3) differential “fitness”: the number of copies of an element that are created in a given time varies, depending on interactions between the features of that element (...) and features of the environment in which it persists [Dennett, 1991, p. 200]

As Dennett’s points out, these condition represent a general characterization of evolution by natural selection, for it says nothing of the matter of the elements concerned – thus the subject of natural selection may be not only organic molecules such as genes, but memes and other entities as well. We can regard the above quoted conditions of evolution as conditions under which it is meaningful to describe the origin of an entity in evolutionary terms, no matter whether the entity has actually evolved according to the evolutionary principles. All the species could have been created by God in six days, but as long as the conditions are satisfied, we can reasonably conceive them as being subjects of evolution. In other words, the difference between the two descriptions lies in their metaphysical assertions, but they amount to the same in consequences, which is what matters most. Therefore even if you regard Dawkins’s *memetics* to be a too far-fetched theory, it nonetheless functions well as a description of how ideas spread and endure in the environment, for the character of the environment and the mechanism of proliferation of ideas among people invite such an interpretation.

The propagation of memes has been enabled by the development of language whereby ideas can be easily transmitted. From the evolutionary stance, the primary goal of a meme is the same as of a virus⁶² – to infest such an environment where it could endure and propagate (cells for viruses, human

⁶²Dennett compares the mechanism of meme propagation to that of genes rather than of viruses. The reason is that viruses are in fact parasitic form of genes that are the real subjects of evolution and natural selection. Genes compete in expansion, and the best way to survive and proliferate is to develop a mechanism of self-defense; hence from

minds for memes). Viruses are obviously bound to their material constitution (they are parasitic DNA or RNA molecules). Memes, on the other hand, are less obviously materialized: in order to transfer from one mind to another, they have to be carried by meme vehicles – pictures, books, utterances, tools etc. Yet once they enter a mind, they do not cease to be physically embodied in some medium – at certain level of description, they are carried by the corresponding brain (more precisely by some particular structures within).

As Dennett emphasizes, the most important consequence of Dawkins's memetics is that "a cultural trait may have evolved in the way it has simply because it is *advantageous to itself*" [Dennett, 1991, p. 204]. This is to say that a meme could have spread among people not only because they believed it was true or that they liked it, but also just because the meme is a good replicator. This kind of Copernican turn in thinking about ideas enables us to explain why certain ideas persist in spite of being deemed as wrong, dangerous, immoral, silly etc. Thus the proliferation of memes like anti-semitism, conspiracy theory, or even suicide can be explained despite of its lack of utility to human beings. Our minds are infested by many memes not because we are convinced of their utility to us, but because our meme-immunological systems are not good enough to get rid of dangerous memes⁶³.

5.2 Mind as a Program

The rather long exposition of memetics was necessary if we are to understand Dennett's final claim that consciousness is shaped by memes, and is, in the end, a complex of memes acting on the hardware of the brain.

The haven all memes depend on reaching is the human mind, but a human mind is itself an artifact created when memes re-structure a human brain in order to make it a better habitat for memes. [Dennett, 1991, p. 207]

the evolutionary point of view, an organism is a large defense mechanism for genes, it is a complicated gene vehicle. I use the analogy with viruses mainly because the above mentioned view of genes as being the elements of natural selection might be unfamiliar to the reader. Besides, the analogy is only to clarify the functioning of memes, and to that extent, viruses serve as well as genes, I believe.

⁶³This may already seem stretched too far – we think we accept idea for the meaning they bear, because we think they are true, for example. But what about tunes from commercials? We do not accept them willingly, yet we can sometimes hardly get rid of them. Besides, "understanding of the meaning," which we think is the reason why we accept or reject an idea, could be just the manifestation that our brain is wired in such a way that it is likely to be infested with that meme. Nevertheless, this elaboration of the consequences of memetics is not so important for our discussion.

How are we to understand the claim that human mind is an artifact created by memes? I will first discuss a less outrageous interpretation which struck me initially. We can think of memes as of elements that make up the content of a pre-existing form of consciousness. As books put in a library make up its content (they make up what the library is like, for example, whether it is a scientific, intellectual's or common reader's library), so memes make up the content of our mind, which nevertheless is of an invariant form. As there has to be first a bookcase, or other place to put books into, so there has to be first consciousness or other medium ready for the uptake of memes. To put it in Aristotelian terms, memes are like matter from which every particular mind is built and consciousness is its form. The essence of mind then would be much closer to the form than to contingent memes that merely "fill in" the form. But the task we have set on is to grasp the essence of mind, or consciousness, and not to find out what the content of a particular mind is. Hence memetics is useful only insofar as it tells us what are the elements which the content of mind consists of. At least two intuitions seem to support this interpretation:

1. In order for memes to enter minds and to work in the above described way, there must be a system which actively uptakes them according to their meaning; hence the system must understand them, at least basically (consciousness is supposed to be such a system, though here considered just as a "white paper," devoid of any "meme inscriptions" at the beginning.
2. Memes themselves do not understand anything, to the contrary, their very existence and "survival" depends on understanding; therefore understanding must be accomplished by something else than memes (consciousness, for example) which consequently cannot exist solely in virtue of memes.

A conclusion can be drawn from the second intuition: there is more to mind than just a huge complex of memes. Both intuitions are Searlean in character, so it is no surprise that Dennett would deny their validity. First of all, he would disapprove the use of Aristotelian framework on consciousness as too a crude structure to be imposed on the natural phenomenon of consciousness that is continuous in character⁶⁴. Moreover, as Dennett holds, consciousness has developed upon subsystems that originally used to serve different purposes, hence the current form of consciousness was not designed

⁶⁴Recall the subsection "Continuity of (the notion of) consciousness" (p. 33).

at one moment but it has rather evolved gradually (by the process of linking up old functions together which resulted into new features).

However plausible the above outlined interpretation may seem, Dennett probably thinks differently, for a few pages after the last quotation, he lays his cards on the table:

Here is the hypothesis I will defend:

Human consciousness is *itself* a huge complex of memes (or more exactly, meme-effects in brains) that can best be understood as the operation of a “*von Neumannesque*” virtual machine *implemented* in the *parallel architecture* of a brain that was not designed for any such activities. The powers of this *virtual machine* vastly enhance the underlying powers of the organic *hardware* on which it runs, but at the same time many of its most curious features, and especially its limitations, can be explained as the byproducts of the *kludges* [i.e. patches, in the programmers’ jargon, that is to say, amendments made *ad hoc* during the debugging of a program] that make possible this curious but effective reuse of an existing organ for novel purposes. [Dennett, 1991, p. 210]

The technical terms in italics are explained in the following pages, but since nowadays (unlike in 1991, when *Consciousness Explained* was published) most of us have at least basic knowledge of how digital computers work, I will discuss them only briefly. The term “von Neumannesque virtual machine” stand for what we know as computer programs – software. Every program is like a machine designed for specific purpose (a word processor, for instance), but since programs consist only of a set of instructions, they have to be implemented in a real machine that actually effectuates the instructions – in that sense they are only virtual. I expect the reader to be familiar with the concept of Turing machine⁶⁵, in the light of which a digital computer (a physical machine) is a Universal Turing machine designed to run other Turing machines. The attribute “von Neumannesque” basically means that the program is designed as a linear succession of steps (executions of instructions follow one after another), and Dennett perhaps finds it important to note because the hardware the program of consciousness runs on works parallelly, i.e. there occur many computations simultaneously. But since Turing showed that computers with parallel and serial (i.e. von Neumann’s)

⁶⁵For a detailed elaboration of the concept of Turing machine see [Turing, 1950, p. 17-19].

architecture are computationally equivalent⁶⁶, it does not matter whether we conceive the program as a set of instruction to be effectuated serially, or as a logical pattern of nodes, their mutual connections and thereon attached weights, which is how a program for a parallel architecture looks like⁶⁷. So Dennett can, in the end, abbreviate his original hypothesis:

Just as you can simulate a parallel brain on a serial von Neumann machine, you can also, in principle, simulate (something like) a von Neumann machine on parallel hardware, and that is just what I am suggesting: conscious human minds are more or less virtual machines implemented – inefficiently – on the parallel hardware that evolution has provided us. [Dennett, 1991, p. 218]

Here we finally have the claim that is so often ascribed to Dennett – that minds are (more or less) programs. A reader familiar with the discussion on Searle's argument will notice the sudden transition from simulation of hardware processes to actual appearance of consciousness. The discussion between Dennett and Searle can be summarized as follows:

Dennett: (1) Consciousness arises upon brain processes. (2) Brain is nothing but a huge and complex neural net whose work consists essentially in transforming the input value into a proper⁶⁸ output value. (3) The transformation of value is a kind of computation which can be done by a Universal Turing machine. (4) Therefore, since consciousness is a product of computation (follows from 1, 2, 3), it is independent on hardware, and can consequently arise on the work of a computer that runs the same program as the biological brain.

Searle: But simulation of a process is not the same as its duplication. As simulation of rainstorm will not make it rain in the laboratory, so simulation of brain processes will not give rise to consciousness because a

⁶⁶That is to say that every computation effectuated by a parallel computer can be effectuated by a serial computer (though perhaps not so fast, in the real time), and vice versa. Dennett himself remarks that "in principle, any parallel machine can be perfectly – if inefficiently – mimicked as a virtual machine on a serial von Neumann machine." [Dennett, 1991, p. 218]

⁶⁷Further information on parallel computing can be found in any literature on connectionism.

⁶⁸The adjective "proper" here means that value which leads to a reaction that is desirable from the brain-bearer's point of view.

computer lacks the causal powers of the biological brain necessary to produce consciousness⁶⁹.

Dennett, however, is careful enough not to say explicitly that minds are just programs, and this is to be pointed out, since many of Dennett's opponents interpret his claims too literally, I think. It is said in the hypothesis (p. 48) that consciousness is a huge complex of memes that "*can best be understood*" as the operation of a program. The hypothesis is introduced by a sentence that expresses it more explicitly:

The level of description and explanation we need is *analogous to* (but not identical with) one of the "software levels" of description of computers: what we need to understand is how human consciousness can be realized in the operation of a *virtual machine* created by memes in the brain. [Dennett, 1991, p. 210]

So the motivation for introducing the concept of a program, as understood in informatics, is that it serves as the best analogy to how memes form consciousness. Dennett's hypothesis surely allows for many interpretations differing in how strictly they take the analogy of consciousness to program, but all the interpretations should take into account the importance of the functionalist attitude to consciousness that underlies the analogy: what really matters about consciousness, indeed what makes consciousness being consciousness, is not what it is made of, but what it does. And since consciousness mainly deals with information and issues commands (which too can be represented in their informational aspect), which can be equally done by a computer, the simulation of consciousness is actually its duplication⁷⁰.

6 Semantics in the Brain

So far I have presented Dennett's explanation of consciousness only by parts without putting much effort to show how these are interconnected. I hope

⁶⁹Searle, in fact, denies the second point in my summary of Dennett's position – he insists that the work of brain cannot be reduced only to the logical structure of information-processing. This is closely related to his conviction that intentionality and semantics are intrinsic features causally dependent on brain processes. As far as I know, Searle has not yet shown an argument supporting his convictions, he considers them to be brute facts that are clear to everyone from the first person perspective.

⁷⁰Unlike with the rainstorm in Searle's example, the simulation of which would have to produce drops falling to the ground in order to be counted as duplication, Dennett thinks there is nothing that a computer could not provide and that is considered to be part of "what consciousness does."

I have made Dennett's view to great extent clear without oversimplifying, or even misinterpreting it too much. I have also tried to point out those parts in his theory that are more or less concerned with the phenomenon of understanding and meaning. I have already sketched some of the points from the discussion on Searle's Chinese Room argument which I think are relevant to Dennett's explanation of consciousness. In this section, I will focus right on the semantic aspect of consciousness, which I regard as its most problematic feature, by bringing together the observations I have made earlier. The reason why I focus on the semantic aspect is that I find it to be the only feature of consciousness that makes the problem of irreducibility of consciousness to physical processes substantiated.

Let me summarize what makes me think we should pay great attention to the origin of understanding and meaning in consciousness as explained by Dennett. (1) As I remarked in the section on heterophenomenology (p. 18), the ability to adopt the intentional stance may be strongly related to the ability to understand, for a part of understanding what one is doing is to know why he or she is doing it, to understand his or her intentions, to see the purpose. In other words, the meaning of an action largely depends on what intentions we ascribe to the agent. On the other hand, we may say we are able to understand someone else's behaviour only because we can ascribe him or her intentions we ourselves are familiar with from the first-personal view on our experience of taking similar actions. (2) Rejecting the Cartesian Theater model of consciousness leads to the denial of a real meaner (remember: there is no Central Meaner), a single, unified thing capable to appreciate meanings, or to recognize symbols as referring to some other things. Of course we understand meanings and we are real, aren't we? Well, it depends on what the pronoun "we" refers to. To (our)⁷¹ selves? But these are rather fictitious characters reconstructed from the information-flow in our brains, hence they aren't good candidates. To (our) bodies? But then it would be nonsensical to claim that bodies understand meanings (unless we are convinced materialists and reductionists). To (our) consciousness? But then again: how much real is consciousness, in Dennett's view? (3) Minds uptake memes (or allow memes to infest them, if you want) not randomly, but for reasons, and these are mainly based on the information the memes convey; hence there must occur some understanding of the memes' message in order for them to enter minds. The understanding is supposed to be consciousness's job, but then there is a threat of circular explanation, for consciousness is

⁷¹It is perhaps circular to use a possessive pronoun if we are uncertain about the subject (Who are "we"? Our X's! Well, whose X's? Ours! And who are "we"? ...), but the grammar does not allow me to express it better.

also supposed to be a complex of memes. (4) Dennett explains away many of the traditional beliefs about the nature of mind as mere illusions. But being a subject of illusion entails understanding, though incorrect (in the case of illusion), what and how things are. So we can be hardly satisfied if anyone explains understanding as a mere illusion. What then is “the real thing” behind understanding?

6.1 Searle and Semantics in the Brain

I do not guarantee that all the above mentioned points are sound, but as long as you grant them some relevance, you will understand my motivation to deal with the problem of semantics in the brain. I should finally clarify what exactly “semantics in the brain” means. The use of the linguistic term “semantics” in philosophy of mind was incited by Searle’s article “Is the Brain’s Mind a Computer Program?” from *Scientific American*, January 1990. There Searle rephrases his original Chinese Room argument in terms of syntax – semantics distinction. The answer to the question in the title of the article is No, for a program merely manipulates symbols, whereas the brain attaches meaning to them. The well-known statement that human minds have (or exhibit) semantics basically means, as far as I understand it, that mind (consciousness) is the domain of sense, the domain where things can appear to be meaningful. Here is what Searle says:

The next axiom [i.e.: “Human minds have mental contents (semantics).”] is just a reminder of the obvious fact that thoughts, understanding and so forth have a mental content. By virtue of their content they can be about objects and states of affairs in the world. If the content involves language, there will be syntax in addition to semantics, but linguistic understanding requires at least a semantic framework. If, for example, I am thinking about the last presidential election, certain words will go through my mind, but the words are about the election only because I attach specific meanings to these words, in accordance with my knowledge of English. [Searle, 1990, p. 21]

Compare it with the following quotation from Searle’s recent book *Mind*:

I have been talking about intentionality and consciousness as if they were independent phenomena, but, of course, many conscious states are intrinsically intentional. My present visual perception, for example, could not be the visual experience it is if it

did not seem to me that I was seeing chairs and tables in my immediate vicinity. This feature, whereby many of my experiences seem to refer to things beyond themselves, is the feature that philosophers have come to label “intentionality.” [Searle, 2005, p. 138]

Thus semantics and intentionality are closely related, because mental states bear meaning (or content) in virtue of reference to other things. Therefore, as Searle holds intentionality to be an intrinsic feature of the brain, so he thinks the same applies to semantics⁷²: semantics, like intentionality, is a real, causal product of the brain. It owes its existence not to merely syntactical algorithm, which is what Dennett’s characterization of consciousness as a program suggests, but to the causal powers of the brain.

6.2 Understanding Memes or Memes Understanding?

Enough of Searle’s view, for the time being. Let’s turn back to the role of understanding in Dennett’s theory. Consciousness, according to the evolutionary argument as expounded in *Consciousness Explained*, has evolved as a by-product of meme infestation of our minds (or brains, if you prefer the physical level of description). But memes themselves are dependent on organisms using sophisticated symbolic system, hence there is no uptake of a meme unless an organism can “read” it, or understand it⁷³. How much is the existence of memes dependent on the evolution of language or other symbolic system? On the one hand, Dennett deals with memes only after he explains how language evolved; he holds language to be the most usual source of meme vehicles, which seems to imply there is a strong relation between the evolution of language and the spread of memes. On the other hand, the examples of meme vehicles Dennett states include also non-symbolic ones – tools, for example. A particular hammer can be said to convey the meme (the idea) of the hammer – a tool for driving nails, breaking up objects etc; an uptake of the meme of the hammer can be mediated not only by language, but simply by observing a hammer in work as well (hence monkeys can get the meme of the hammer too, though they have not mastered language).

So what is the relation between language and memes which give rise to consciousness? I propose the following: every uptake of a meme requires

⁷²Cf. [Searle, 1990, p. 25]

⁷³Dennett himself proposes the example of a pigeon in a city that is exposed to the same amount of meme vehicles as we are (advertisement, signs, labels etc.) but that nevertheless cannot accept memes because it does not possess the appropriate subsystem for parsing symbols into their meanings. See [Dennett, 1991, p. 204]

certain degree of understanding (or misunderstanding) of the meaning the meme conveys; language then is an exquisite tool designed specially for conveying meanings and therefore it provides abundance of meme vehicles⁷⁴. Consequently, language significantly enhances the meme flow between the language users and thereby it enables sophisticated meme complexes to arise in human minds (and consequently to form consciousness). If this is indeed what follows from Dennett's exposition of consciousness as operation of meme complexes, then two remarks must be made

1. Since consciousness is defined as "a huge complex of memes" and since the domain of memes are human minds⁷⁵, the definition of consciousness rests with the definition of mind, which is rather obscure, given the fact that Dennett uses terms "consciousness" and "mind" as synonyms.
2. Since uptake of memes seems to be related to understanding the meaning they convey, and given that such understanding is a non-trivial, higher-order cognitive function, it becomes very important to explain how understanding arises on brain processes, at least insofar as we are interested in the explanation of consciousness.

As to the first point, Dennett's argumentation is not so circular as it may seem from my interpretation. Consider one more time an already quoted passage: "The haven all memes depend on reaching is the human mind, but a human mind is itself an artifact created when memes restructure a human brain in order to make it a better habitat for memes" [Dennett, 1991, p. 207]. It follows from this sentence that the actual domain of memes are brains; minds are brains' products and we speak about memes infesting minds, rather than brains, simply because it is a more apt level of description. More specifically, mind is a product of those structures in the brain that are physical embodiments of memes. But then again, if minds are really only products of earlier meme infestation, who or what does the understanding that we deem is necessary for a meme to be accepted? This question partly proclaims the belief in a "man in charge," an audience in the CT, and Dennett, anticipating as always many of intuitive objections, is ready with his answer:

⁷⁴In the evolutionary terms, it wasn't until the evolution of language that memes could spread so easily, since even though they occasionally emerged (as an idea of a tool, for example) in some ape's mind, they didn't have the right means of transport to infiltrate another ape's mind. The evolution of a symbolic system caused a true revolution both in the spread and in the sophistication of memes; yet they can be said to have existed earlier in ape's minds, sometimes transduced by way of an illustrative example, but mostly becoming extinct with the death of the ape.

⁷⁵Cf. [Dennett, 1991, pp. 206-207]

But if it is true that human minds are themselves to a very great degree the creations of memes, then we cannot sustain the polarity of vision with which we started: it cannot be “memes versus us,” because earlier infestations of memes have already played a major role in determining *who or what we are*. [Dennett, 1991, p. 207]

It is true that if we adopt the stance proposed by Dennett, we cannot ask “who” decides whether a meme will be accepted or not, but we can still ask who or what does the understanding, which does not entail the belief in a Cartesian Self. The only possible answer seems to be “The brain does, of course.” A natural reply would be “And how does it do it?” This is, I think, the right question to ask if we want to understand how all the parts of Dennett’s theory come together. Unfortunately, it is also one of the most difficult questions to which Dennett does not give a clear and straightforward answer. I will nevertheless try to deduce the answer from what has already been said on Dennett’s account.

6.3 Representation in the Brain

Let me first make clear what is to be explained, i.e. what the understanding amounts to. The verb “to understand” is used in many different meanings, but let’s stick to the standard account that to understand the meaning of (1) an assertive sentence is to know under which conditions the sentence is true (truth-conditional semantics); (2) a word (a noun) is to know what it refers to; (3) an action is to know its causes (physical causes, intentions, etc.) or purposes. This is perhaps not very much helpful either, for it postpones the explanation of understanding to that of knowing (to understand is to know. . .). On the other hand, we won’t find a better account of understanding in linguistic semantics, for semantics as a part of the theory of language is concerned with relations between language expressions and meanings, and ultimately appeals to the faculty representing things as standing for other things⁷⁶. If “reference” is the basic relation of semantics, and if it is related

⁷⁶A semanticist might say: “The meaning of an expression is determined, according to the principle of compositionality, by the meaning of its constituents. The elementary expressions that bear meanings are words, and the meaning of a word (a noun) is its reference (extension). Ultimately, a word refers to some other thing in virtue of human faculty of taking the word as a symbol representing some other thing.” I am well aware that this is a crudely simplified exposition of a semantic theory (it omits many other things that are related to meaning of an expression, such as context, sense (Sinn) of a word, the meaning of adjectives and verbs, etc.), but I hope it shows why semantics cannot offer us the right answer.

to the faculty of representing things, let's see how it originates in the brain. What is the ground for the property of a mental state, and hence of the corresponding neural activation, that it represents something?

For instance, it seems that the only sort of facts that could explain a particular neural tract's "caring about" color would be facts about its idiosyncratic connections, however indirect, to the cone cells in the retina that are maximally sensitive to different frequencies of light. Once such a functional identity was established, these connections might be cut (as they are in someone blinded in adulthood) without (total) loss of the power of the specialists to represent (or in some other way "care about") color, but without such causal connections in the first place, it is hard to see what could give specialists a content-specific role. It seems then that the cortex is (largely) composed of elements whose more or less fixed representational powers are the result of their functional location on the overall network. [Dennett, 1991, p. 272]

So, if a particular neural circuit gets established after you have been exposed to a number of stimuli of the same kind – for example an image of an animal previously unknown to you –, and is subsequently active whenever you perceive the stimulus, then that neural circuit becomes representing the animal. Dennett mentions, though in rather an obscure way, a causal theory of reference in a footnote attached to the quoted passage, which again emphasizes the importance of the causal connection between the representing and the represented⁷⁷.

The last quoted sentence expresses the most important feature of representation. The representational power of a bunch of neurons lies not only in its activity in the presence of this or that stimulus, but also in its relation to other representations, in its functional role, which cannot be completely detached from the functional roles of other representations. For example, what makes some neurons represent a cat is not only that they are active whenever there is a cat perceived, but also that they are not active when a dog or a chair is perceived. If it is possible for a symbol to acquire meaning solely by the way it is used, why should it not be possible for neurons? As

⁷⁷But what would count as "the initial baptism," which is what establishes the referential relation between a noun and its reference? Possibly the re-wiring of some neurons in precisely that way that they become active whenever the stimulus they represent is present. But then again you have to take into account that neurons in the brain are being constantly re-wired and yet a small change in the wiring of the neurons involved in particular representation does not lead to a total loss of the faculty to respond to the represented stimulus.

far as I understand the last quoted sentence, it expresses the same idea as Wittgenstein's "the meaning of a word is its use in the language." Analogically, we could say that "the meaning of a pattern of neural activity is its use in the interaction with the environment." I am not sure if I have made the idea comprehensible but I can't do any better than that.

Let's assume we have a hunch as to how representation is done at the brain level. Now we have to turn to the question how it comes that consciousness (whatever it is) is able to appreciate and deliberate on meanings. It is hard, of course, to say what exactly is this appreciation of meaning, and Dennett doesn't forget to point it out. Let's take an example. Consider a quote by E. M. Forster, that is, by the way, the motto of Dennett's explanation of the relation between consciousness and language: "How do I know what I think until I see what I say?" Perhaps you have just started pondering about the meaning of the quote. The appreciation of the meaning, judging whether it is witty, false or nonsensical is a very complex process, indeed so complex and intense that it is hard to believe it can be explained in terms of information-processing as done by neurons. Obviously, by introducing the expression "appreciation of meaning," I have resorted to more or less ineffable subjective experience, which is unjustifiable (or at least unscientific) step according to Dennett. All I can appeal to is that understanding a sentence such as Forster's *seems to me* to be clear and perfectly rational, and yet I cannot even tell the rules according to which I derived the meaning of the sentence. Nobody would expect me to be able to tell the rules of the inference of meaning simply by reflecting on the very process of understanding, for then it would be rather an easy task for semanticists to find out the rules and principles. Most likely, there are no such rules and principles involved in the process of understanding, given that the brain follows flexible patterns rather than strict rules⁷⁸. But how come that understanding the sentence entails, phenomenologically, a clear representation of what the sentence means, its possible paraphrases, the appreciation of its wittiness, etc.? How come that I feel I know not only the reference of each word, but the sense (Sinn) of it; and how come that I find the sense of each word quite clearly delineated, as if I were in direct contact with some platonic realm of sense?

Most people do not see how these questions could be answered in terms of a program blindly following some prescribed rules. Precisely this intuition is exploited by Searle's Chinese Room argument; and Dennett reasonably replies that the fact that it is hard to *imagine* the appreciation of meaning to be a result of mechanical information-processing nonetheless proves nothing,

⁷⁸The rules for derivation of truth-conditions from the constitutive expressions would be at best good approximations of what the brain does at subsymbolic level.

for the mere difficulty of imagination does not necessarily lead to impossibility⁷⁹. I agree that this powerful “Searlean” intuition may be wrong, but it is then up to Dennett to help us abandon the intuition by clearly showing how understanding works. He tries, of course, but I am not quite certain whether to deem his attempt successfully, for I still feel I lack the desirable insight into the subject matter. This is not to say he does *not* explain it at all: I cannot exclude the possibility of me being too dull to be able to grasp the transition from information-processing to the experience of understanding. Let me therefore get back to the text and arguments.

6.4 Dennett and the Chinese Room

Considering what Dennett would say, if he had to face Searle’s Chinese Room argument, might help us to find a new idea about the relation between information-processing and understanding. The point of the Chinese Room argument is basically that the room produces sentences indistinguishable as to the meaning to those of a native speaker, and yet the symbol-shuffler understands nothing. From the replies Searle listed in the version of the argument from the article “Minds, Brains and Programs,”⁸⁰ Dennett would probably side with the system reply⁸¹ which says it is the room as a whole that understands Chinese, though none of its components does (Searle, or the man handling Chinese symbols, notwithstanding)⁸². The system reply perfectly fits into his model of consciousness as constituted by many parallelly flowing “narratives” and his reluctance to look for a single, unified source of mental contents. Who does understand meanings then, according to Dennett? Generally, it is the program of consciousness the code of which is written by memes; concretely, it is this or that brain. According to Searle, however, the claim that the understanding is done by a program is already wrong, for programs are purely syntactical, and syntax is neither constitutive of nor sufficient for semantics:

Axiom 3. Syntax by itself is neither constitutive nor sufficient for semantics. At one level this principle is true by definition.

⁷⁹Cf. [Dennett, 1991, p. 282]

⁸⁰Searle, J. (1980). “Minds, Brains and Programs” *Behavioral and Brain Sciences*, 3.

⁸¹Cf. [Dennett, 1991, p. 439]

⁸²If we go one step further and consider Searle’s reply, saying that we can imagine the man inside as having internalized the whole process of symbol-shuffling and yet understanding nothing, Dennett would perhaps disagree with the claim that the man still understands nothing, since from his functionalist attitude Dennett cannot do otherwise. If it can engage in a dialogue like a Chinese, if it talks like a Chinese, then it understands Chinese.

One might, of course, define the terms syntax and semantics differently. The point is that there is a distinction between formal elements, which have no intrinsic meaning or content, and those phenomena that have intrinsic content. [Searle, 1990, p. 21]

In the same issue of the journal *Scientific American*, where the above quoted Searle's article appeared, Paul and Patricia Churchlands doubt the certainty of Searle's third axiom by presenting an argument from analogy. They propose to consider an argument that is formally identical to Searle's, yet its conclusion is false:

Axiom 1 *Electricity and magnetism are forces.*

Axiom 2 *The essential property of light is luminance.*

Axiom 3 *Forces by themselves are neither constitutive of nor sufficient for luminance.*

Conclusion *Electricity and magnetism are neither constitutive of nor sufficient for light.* [the Churchlands, 1990, p. 29]⁸³

Searle's argument from syntax – semantics distinction is backed up by the Chinese Room thought experiment⁸⁴, where all the operations executed by the man inside proceed so slowly that it makes our intuition deny the presence of any understanding at all. The Churchlands likens it to a thought experiment with a man pumping a bar magnet in a dark room. Intuitively, we would deny the presence of light in the room as well, but the light is there, in fact, though of such a long wavelength that human retinas cannot respond to it. It is then only a matter of frequency whether a pumping magnet will produce light or not, and so the Churchlands argue that, provided the analogy between semantics and luminance holds, execution of the right program at some speed could yield consciousness. In order to illustrate the importance of speed, they refer to connectionism, a branch of AI that does research on computers which process input values parallelly (unlike digital computers), and whose architecture is similar to the brain's.

Searle is not, of course, convinced whether the analogy between semantics and luminance holds, and since the intuition seems to be on his side, the burden of proof lies on those who think semantics, or consciousness in general,

⁸³For a corresponding formulation of Searle's argument see [Searle, 1990, p. 21].

⁸⁴As J. Moyal points out, it is important to distinguish between the argument and its illustration by means of the Chinese Room thought experiment, see [Moyal, 2003, p. 218].

can emerge if the right program is executed. Here again we see the need for a clearer demonstration of how a program can yield understanding, and thence consciousness. I am ready to accept that the *seemingly real* nature of meaning and its understanding is only an illusion, but I have not been, unfortunately, shown how the illusion works.

On what conditions, however, can it be shown? Let me consider again the methodological question that I first put forward in the subsection “Nature of the Explanation,” this time rephrased as: What would count as an explanation of semantics? Perhaps, I have unwillingly expected an explanation that would tell me why understanding *feels* like this or that, and why the sense *seems to be* like this or that. That may be the origin of the difficulties that I have with Dennett’s account of consciousness. However, so much have I understood that I recognize two answers to the question “Can it be shown how the first-personal aspect of understanding results from the third-personal aspect of representation in the brain?” Yes, insofar as we focus mainly on how understanding manifests itself in our intelligent behaviour, and how meaning originates in our interaction with the environment. No, if we insist on the explanation of the nearly mysterious phenomenological qualities of understanding.

7 Conclusion

I intended this paper to roughly outline Dennett’s account of consciousness as expounded in his book *Consciousness Explained*. This alone is rather a difficult thing to do, since Dennett’s style of argumentation prefers examples, stories and thought experiments against clearly stated assertive propositions that would follow from some previously mentioned premises. Dennett’s greatest narrative power lies in his ability to make the reader realize that what he or she took to be a simple fact may just be an illusion. Dennett puts in doubt many of our deeply rooted intuitions about the nature of mind but the replacement he offers is seldom as simple as the argument against the intuition, let alone the intuition itself. This may lead to a kind of frustration – the reader may feel robbed of some precious knowledge (e.g. what consciousness is) without obtaining an adequate compensation. This is rather a common situation when things and concepts get into philosopher’s hands, but Dennett undermines that what matters most to us (“the greatest conceivable thing”) – our consciousness. All this brings about specific difficulties when interpreting and understanding Dennett’s works. For this reason I chose first to discuss the different parts and arguments of his theory – it is easier to understand the partial arguments and interpret them separately, than to see their

the connections between them and incorporate them in an all-encompassing interpretation. I hope I have succeed at presenting the parts; it is up to the reader whether I have succeed in my attempt to show the interconnectedness of the parts by discussing the problem of semantics in the brain.

The actual conclusion of this paper depends to a certain extent on the reader. At the end, I addressed the question whether Dennett's theory explains satisfactorily a particular feature of consciousness – understanding. I have tried to provide the reader with enough evidence to be able to find his own answer. For my part, I regard Dennett's theory as unsatisfactory, insofar as the phenomenon of semantics in the brain is concerned, but this stems from the fact that I am just not convinced that his theory covers all the features of the phenomenon of understanding⁸⁵. On the other hand, Dennett's aim is not to provide a complete theory explaining consciousness, but rather to show that such a theory is possible at all:

My main task in this book is philosophical: to show how a genuinely explanatory theory of consciousness *could* be constructed out of these parts, not to provide – and confirm – such a theory in all its details. [Dennett, 1991, p. 256]

Dennett succeeds in this, I think, but I still find the title of the book *Consciousness Explained* promising more than what can actually be found within. *Consciousness Demystified* would be perhaps a more adequate title, and demystifying consciousness alone is a creditable result of Dennett's effort. It is actually the first step to be taken on the way to explanation of consciousness.

⁸⁵To put it simply, I am by no means convinced that there is something wrong in his theory, I just may not understand it fully.

References

- [the Churchlands, 1990] Churchlands, P. and P. (1990). "Could a Machine Think?" *Scientific American*, January 1990, pp. 32-37
- [Dawkins, 1976] Dawkins, R. (1976). "Selfish Genes and Selfish Memes," in D. R. Hofstadter and D. C. Dennett eds. *The Mind's I* (1981). New York: Basic Books. pp. 124-144.
- [Dennett, 1991] Dennett, D. C. (1991). *Consciousness Explained*. New York: Little Brown.
- [Dennett, 1984] Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: The MIT Press.
- [Dennett, 2001] Dennett, D.C. (2001) *The Fantasy of First-Person Science*. 1 March, 2001. Center for Cognitive Studies, Tufts University. [online] [cited August 21, 2006] Available from Internet: <<http://ase.tufts.edu/cogstud/papers/chalmersdeb3dft.htm>>
- [Dennett, 1997] Dennett, D. C. (1997). *Kinds of Minds*. London: Phoenix.
- [Dennett, 2005] Dennett, D. C. (2005). *Sweet Dreams: philosophical obstacles to a science of consciousness*. Cambridge: The MIT Press.
- [Flanagan, 1992] Flanagan, O. (1992). *Consciousness Reconsidered*. Cambridge: The MIT Press
- [Haugeland, 1998] Haugeland, J. (1998). "Understanding: Dennett and Searle," in J. Haugeland, *Having Thought: Essays in the Metaphysics of Mind* Cambridge: Harvard University Press.
- [Levin, 2004] Levin, J. (2004) "Functionalism," in *Stanford Encyclopedia of Philosophy* [online]. 2004 [cited August 21, 2006] Available from Internet: <<http://plato.stanford.edu/entries/functionism/>>
- [Moural, 2003] Moural, J. (2003) "Searles Chinese Room Argument and its Relatives," in B. Smith ed. *John Searle (Contemporary Philosophy in Focus)*. Cambridge: Cambridge University Press, pp. 214-260.
- [Nagel, 1974] Nagel, T. (1974). "What Is It Like to Be a Bat?" *Philosophical Review*, 83, pp. 435-450.
- [Searle, 1990] Searle, J. (1990). "Is the Brain's Mind a Computer Program?" *Scientific American*, January 1990. pp. 26-31.

- [Searle, 2005] Searle, J. (2005). *Mind: A Brief Introduction (Fundamentals of Philosophy)*. New York: Oxford University Press.
- [Turing, 1950] Turing, A. (1950). "Computing Machinery and Intelligence," *Mind*, 59, pp. 433-360.