Univerzita Karlova v Praze

Filozofická fakulta
Ústav filozofie a religionistiky

Diplomová práce

Martin Vraný

Speaking the Mind, Minding the Language

Praha, 2010

vedoucí práce: Prof. RNDr. Jaroslav Peregrin, CSc.

Poděkování

Děkuji vedoucímu práce J. Peregrinovi za přednášky, v nichž mi ukázal, že filosofické otázky je možné přesně formulovat i bez pomoci odborných výrazů. Děkuji J. Palkoskovi za to, že mě učil důslednosti v analytickém rozvažování. V neposlední řadě děkuji P. Koubovi za připomenutí, že filosofická práce nespočívá v nezpochybnitelné argumentaci.

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

V Praze dne 3. ledna 2010

Martin Vraný

Anotace

Práce ukazuje, že k problému vztahu těla a mysli, a konkrétně k otázce vysvětlitelnosti vědomí přírodními vědami, je třeba přistupovat skrze zkoumání jazyka a významu. Za jádro problému vztahu těla a mysli je označena kantovská transcendentální jednota apercepce a je zdůrazněn rozdíl mezi empirickým a transcendentálním vědomím. Předpokládá se, že empirické vědomí je uspokojivě vysvětlitelné pomocí teorie myšlenek vyššího řádu. Následuje rozbor různých aspektů významu, intencionality a užití jazyka, které podporují závěr, že podmínky možnosti býti mluvčím jazyka s sebou nesou transcendentální podmínky vědomí. O jazyku lze tak říci, že konstituuje vědomí nejen v tom smyslu, že kritéria připsání vědomí něčemu (někomu) jsou ve své podstatě jazyková, ale též ve smyslu, že vědomí se objevuje se schopností mluvit.

klíčová slova: vědomí, jazyk, význam, pojmy, intencionalita, transcendentální sebevědomí

Abstract

The thesis proposes to address the mind-body problem, and specifically the question of scientific explanation of consciousness, in terms of language and meaning. First, the core of the mind-problem is identified with Kant's transcendental unity of apperception and the distinction between empirical and transcendental consciousness is emphasized. Empirical consciousness, as consciousness *of* something, is assumed to be best approached by a higher-order theory of consciousness. Then various aspects of meaning, intentionality and language in use are discussed to prepare ground for the conclusion that transcendental conditions of consciousness are entailed by conditions of being a genuine speaker of language. Thus language can be said to be constitutive consciousness not only in the sense that the behavioural criteria for attributing consciousness are essentially linguistic, but also in the sense that consciousness comes with the ability to speak.

keywords: consciousness, language, meaning, concepts, intentionality, transcendental selfconsciousness, mind-body problem

Contents

1 Introduction

The mind-body problem seems to be the center of the philosophy of mind around which most of its issues revolve. Despite its long tradition, the philosophy of mind has faced a great challenge caused by the practical and explanatory success of natural sciences. Various aspects of the mind have undergone scientific scrutiny - sometimes to considerable success, other times to almost no success at all. Results and methods of empirical research have exerted pressure on philosophers and theorists to specify what the mind actually is, what is to be explained. Ingenious metaphysical theories have been developed either to conform to both "hard" scientific facts and our reflective investigation of the mind, or to show, by speculation or conceptual analysis of the mind, that contemporary scientific knowledge is not sufficient for a viable explanation of the mind. Thus today the philosophy of mind is enriched with concepts like supervenience, property dualism, physicalism, emergentism, reductionism etc. Perhaps too much of philosophical effort has been devoted to conceptual analyses to the desired effect that consciousness is in principle inexplicable in scientific terms. I am convinced, and I will not argue further for this belief as I find it a fundamental one, that consciousness is a natural phenomenon. It follows then that it is itself an empirical matter whether the mind-body relation will be satisfactorily explained by science. We should not claim now that it *cannot* be explained, even if we were quite confident that contemporary scientific knowledge does not comprise concepts appropriate for the explanation of some aspects of the mind, for the scientific paradigm is ready to change with any groundbreaking discovery. And if it is to be found out that some crucial question about the mind cannot be answered due to practical or physical limits then again it is not a conceptual but an empirical matter that consciousness *cannot* be explained scientifically.

1.1 Outline

In general, the point of this thesis is to argue that in the attempt to naturalize consciousness the focus should be on the use of language. The fundamental idea behind the point, which is not original, is that language is constitutive of consciousness. What follows is an attempt to clarify the relation between language and consciousness so that the idea of constitution makes better sense. I begin with arguing against an influential tendency in the philosophy of mind which takes it as an essential feature of consciousness that there be a certain feel to conscious experience. I argue that appealing to feels or what-it-is-likeness is incoherent and so it cannot support the claim that consciousness is irreducible to brain processes or anything else that science

can describe. I do not think, however, that rejecting the notion of what-it-is-likeness makes consciousness easier to explain. I identify the core of the so-called Hard problem of the mind-body relation with Kant's transcendental unity of apperception, which, despite its structural and functional specification, is still hard to account for. Subsequently, I argue that the conceptual, which is the 'medium' of judgements, is inextricably linked to the linguistic, and hence that the best way to learn about the possible foundation of the transcendental unity of apperception is by looking at the linguistic.

I continue with some remarks on the nature of the conceptual and I argue that, despite its tight link to the linguistic, it is a concept worth keeping in the philosophical vocabulary. Afterwards, I turn to the discussion of the principle of compositionality. Though the principle originally applies to meanings of linguistic expressions, I argue that an analogous principle has to be supposed to work in human understanding if productivity of our thinking is to be explained.

In section 4 I finally discuss three different approaches to the nature of intentionality and meaning which, according to J. Haugeland, cover for all positive accounts of these philosophical concepts. Haugeland labels the approaches neo-cartesianism, neo-behaviourism and neo-pragmatism. Works of major proponents of each of the approaches are then discussed in more detail in order to look for observations that might contribute to our understanding of the relation between language and consciousness. A parallel result is that one's semantics strongly constrains one's theory of mind and vice versa, which again supports the idea that the attempt to explain consciousness should begin with explaining the linguistic capacity. The selected proponents are J. Fodor, D. Dennett and R. Brandom respectively. I do not come up with a synthetic interpretation that would unite bits and pieces from each of the theories. I arrive at the conclusion that Fodor's approach is likely to be misdirected, though I appreciate that his theory makes it clearer than the others, due to strong ontological commitments entailed in computationalism, what is to be done to explain rational thinking or human mindedness[1] - find the right program the execution of which produces rational sequence of thoughts. Dennett's theory contributes mainly to the explanation of empirical consciousness, and little can be inferred about what Kant would call transcendental consciousness from his works. Unsurprisingly, it is Brandom's philosophy, which explicitly elaborates Kantian themes, that may offer some insights into the linguistic foundation of the transcendental unity of apper-

[1]Fodor carefully avoids speaking of consciousness. His explanatory effort is aimed 'only' at rational thinking, which he understands as a sequence of various propositional attitudes to mental contents.

ception. That is the topic of the last section, where it is concluded that transcendental conditions of consciousness are entailed in the conditions of being a speaker.

To make it clear, the thesis does not design a theory of consciousness. Its main purpose is to support the idea (and specify its sense) that language is constitutive of consciousness. No doubt, language may be constitutive of consciousness in other senses as well. I have chosen the transcendental view because I find it the right way to approach the Hard problem; even though I realize that by plunging into the Kantian paradigm I run a greater risk of misunderstanding than in others.

1.2 Some ontological remarks about the mind

Although I believe that the subsequent treatment of the topic is not bound to a specific metaphysics of mind, some of the claims about the mind that I take for granted in this thesis betray commitment to what could be considered an Aristotelian. According to this view, the soul (the mind)[2] is the form of the body in much the same way as a house-like structure is the form that makes a mass of bricks, mortar and beams a house.[3] Thus strictly speaking, human being is a compound of the body and the invigorating mind; and these are inseparable from one another as a house-shape is inseparable from the material which fills it out. However, for the specifc purpose of elucidation of the mind-body relation I can afford to take such expressions at face value, leaving aside most of the problems of hylomorphic metaphysics, and emphasise only those threads of the Aristotelian framework that are congenial with my view.

First, the framework provides a useful direction how to look at the problem of the unity of the mind and the body. The mind and the body can be regarded as identical or different in the same sense as the wax and its shape. If platonic dualism is considered to be the right metaphysics, one could interpret forms as entities that are somehow conferred to the matter in the world of appearances and whose existence is independent of the matter. Thus the mind would be a things distinct from the body, as the Cartesian dualism holds. If materialism is considered to be the right metaphysics, forms of complex things, such as human beings, could be interpreted as reducible to structure of the matter which consists of nothing but elementary particles

[2]While Aristotle conceived of the soul as the unity of various faculties, i.e. nutrition, perception and mind, we may substitute "the mind" (in the contemporary sense) for "the soul" for the purpose of elucidating a specific view of the mind-body relation. Thus the conception is Aristotelian in a very relaxed sense.

[3]The parallel as well as the general overview is taken from [Shields(2000)].

(whose 'form' is fundamental) and their relations (presumably spatiotemporal). Thus minds would be results of complex material organization. The crucial advantage of the Aristotelian framework is recognizing the mind as that what *essentially* makes some material body a human being - and it is a further question whether (or in what sense) it implies that minds must be distinct from bodies or not.

Second, the sense in which the mind is the form of the body of a human being is functional. As a mass of bricks and beams is a house only if it serves the function of a shelter appropriate for living in, so the body is a living human being only if it thinks, perceives, acts rationally, etc. Thus the mind can be described in terms of what it does, rather than how its working is realized. Admittedly, this is an expression of functionalism in its *broad* sense; and I take it as another fundamental belief for which I will not argue in the thesis. A corollary is that minds can be conceived of only as *embodied* if rational agency is deemed to be part of the functional specification. So the famous arguments that computers cannot think or be conscious, which are to support claims about some sort of irreducibility of the mind to the body, are aimed, so to speak, at a too easy target, for ordinary computers lack *interest* in the world upon which they would like to act.[4] Thus for something to exhibit 'mindedness', it must not only be able to exhibit certain cognitive faculties on demand (that is: interact with environment), but it must also have a goal that it could try to achieve by rational behaviour. Perhaps, the range of goals upon which someone acts constrains what it is to be a human with equal strength as the range of faculties used in achieving the goals.

Now, one could argue that endorsing functionalism is a controversial step, because it greatly simplifies the whole issue of consciousness and intentionality by disregarding some purportedly essential questions. I don't think this hypothetical objection is quite right, for it is usually aimed at functionalism in the narrow sense, which is a doctrine about the nature of *mental states*. Though functionalism in the broad sense, as a doctrine about the nature of the mind, is still a view that some theorists would find controversial, it is a view compatible with the scientific paradigm, unlike the Cartesian dualism, for example. Now, since it is generally agreed by most theorists that minds are natural phenomena and that scientific approach is the most fruitful one in explaining nature, it ought to be supposed, as a null hypothesis, that the functionalist view of the mind is adequate. Of course, an argument against functionalism in the broad sense is likely to be based on analyses of the scientific paradigm as such, with the conclusion that it lacks some concepts necessary for an adequate explanation of the mind. But as I have argued

[4] Cf. [Searle(1990)].

above, I do not think such analyses are sound, because if minds are natural phenomenam then the possibility of their explanation is not a conceptual matter.

Finally, I don't think that the Aristotelian threads I have chosen to follow are incompatible with the Kantian theme adopted below. Roughly speaking, the Aristotelian view provides a framework for empirical conception of the mind-body relation, whereas the Kantian paradigm shows the preconditions of 'mindedness', which may, or may not, invite an empirical explanation. Hopefully, it does make sense to detach the Aristotelian view of the mind from Aristotle's metaphysics, and also to follow Kant's transcendental considerations about consciousness without endorsing the doctrine of transcendental idealism.

2 Shifting the Hard problem

If the mind-body problem is really central to the contemporary studies of the mind, what is the essence of the problem? Why do we still talk about the *mind-body* problem if there are but a few philosophers willing to conceive of the mind and the body as of two different substances? Although many people are inclined to think of the body and the mind as a special kind of unity and hence want to abandon the Cartesian framework, in which the core of the problem is intersubstantial causation, we still have no generally accepted idea how to conceive the unity. However, we should not conclude too hastily that the gap between the body and the mind is merely conceptual. Although most of the arguments against the prospect of science to explain the mind make use of the apparent incompatibility of the concepts used for describing the mental and the physical, they often go one step further, arguing that such conceptual incompatibility indicates some sort of metaphysical difference. David Chalmers presents the argument from incompatibility (also known as the explanatory argument) as follows:

(1) Physical accounts explain at most structure and function.

(2) Explaining structure and function does not suffice to explain consciousness.

---

(3) No physical account can explain consciousness.[5]

[5] [Chalmers(2003), p. 103].

While we may accept the first premise without serious objections, we must ask on what ground the second one is justified. According to Chalmers, who follows an idea originally expressed by Thomas Nagel in his influential paper "What Is It Like to Be a Bat?", structure and function alone cannot explain experience, which is a defining feature of consciousness.[6] Experience is, according to Chalmers, what makes the problem of the scientific account of consciousness really hard:

> The really hard problem of consciousness is the problem of *experience*. When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism. This subjective aspect is experience. [Chalmers(1995), p. 3]

In this section, I would like to 1) specify what Chalmers's (and Nagel's respectively) notion of experience amounts to; 2) show that this notion of experience is either too vague or incoherent for the demand of its scientific explanation to be reasonable; 3) argue that a meaningful explanation of consciousness need not account for the intuition behind that notion of what-it-is-likeness; 4) suggest that the hard problem be shifted from experience to meaning.

2.1 The Hard problem

In order not to just change names for the problematic domain, Chalmers tries to pin down what he means by experience:

> Human beings have subjective experience: there is something it is like to be them. We can say that a being is conscious in this sense - or is phenomenally conscious, as it is sometimes put - when there is something it is like to be that being.[7] [Chalmers(2003), p. 103]

[6] It is no coincidence that consciousness attracts most of the attention when it comes to arguments about the possibility of scientific explanation of the mind. Consciousness is an aspect of human mind that makes the mind-body problem hard. No other aspect of mind, such as memory, perception etc., is considered irreducible. Thus the mind-body problem seems to 'reduce' to the consciousness-body problem.

[7] Other versions of the definition ([Nagel(1974)], [Chalmers(1995)]) read 'organism' instead of 'being'. I think it is an unjustified restriction since the what-it-is-likeness is presented as both a sufficient and necessary condition of consciousness, hence it should pick out conscious beings even under universal quantification. Why not limit the definition even further, for example: "a higher mammal is conscious if ..."? Clearly, the restriction to organisms only is motivated by a tacit assumption that anorganic compounds do not

For the sake of analytical clarity, it is usually accompanied by a definition of the conscious mental state, because consciousness is predicated not only to beings but to mental states as well. Analogically, a mental state is conscious iff there is something it is like to be in that state. This *something* Chalmers and Nagel refer to is technically called quale or phenomenal property. Since Chalmers evidently wants to employ the criterion of what-it-is-likeness in the original Nagel's sense, I will refer in the following discussion to Nagel's statements mainly. Nagel's formulation is:

> [A]n organism has conscious mental states if and only if there is something that it is like to *be* that organism - something it is like *for* the organism. [Nagel(1974)]

The ontological commitment implied by the quoted definition, which is not accidental as there are many other expressions in Nagel's and Chalmers's articles bearing the same commitment, demands to treat qualia as real things, or at least to adopt a factualist stance about phenomenal properties. Thus the logical form of the quoted conditional is most naturally read as: $\forall x$ ($\exists y$ (Q(x,y)) $\leftrightarrow$ C(x)); x is 'C'onscious *iff* there is y that is 'Q'uale of x - its what-it-is-likeness.

There is nothing strange in what-it-is-likeness being taken as a relational predicate, since qualia are from the beginning construed as felt qualities *of* something. According to this logical form, unless 'Q' is a reflexive relation (it cannot be, intuitively, a symmetrical relation), qualia cannot be conscious, but that does not necessarily violate our intuitions, for qualia are *felt* objects, and though we may be conscious *of* them, they themselves are not conscious. To conclude this analytical nit-picking, let's reserve the term 'qualia' to y's and let the predicate 'Q' be called what-it-is-likeness.

Since we know (empirically) that there are conscious organisms, we have invited qualia to our metaphysical garden by this conditional. To be is to be the value of a (bound) variable. It will be first argued that this introduction of qualia among basic metaphysical categories is highly problematic due to uncertain criteria of identity. Afterwards, it will be explored whether postulating qualia is really necessary for the explanatory argument to work, or more generally, whether Chalmers's intuition turns out to be void if we hold a non-factualist position about phenomenal properties.

---

sense anything and hence there cannot by anything it is like to be them. I think this is an unjustified assumption.

2.1.1 Identifying qualia

Why is it untenable that qualia are to be considered real? Because we do not have any criteria of their identity that would make them what they are supposed to be. What are qualia supposed to be? D. C. Dennett, who steadfastly argues against the usage of the notion in the philosophy of mind, recognizes four characteristic features of qualia, based on works of their proponents.[8] They are: 1) ineffable, 2) intrinsic, 3) private and 4) directly or immediately apprehensible in consciousness. This analysis is in accord with Nagel's usage, for his argument to the conclusion that we may never know what it is like to be a bat makes use of all of them.[9] I find it unnecessary to elaborate more the concept of quale, because for the sake of my argument, the present outline is enough; besides, as Dennett observes, there is no clearer definition of the concept and its sense is usually evoked by examples.[10]

Why do we need criteria of identity in order to consider a thing real? It is a reasonable methodological regulative which says that one may introduce a new type of object to a theory[11] if she can provide it with such criteria (which ought to be stated in terms of the yet unextended theory). Thus mathematicians may introduce fractions as $\frac{a}{b} =_{df} \frac{c}{d} \leftrightarrow b \cdot c = a \cdot d$, where there are already familiar expressions in the *definiens*. Similarly, physicists may introduce black holes, electrons or strings, provided they say how these objects manifest themselves in a way that can be observed. Identity criteria delimit a thing out of the whole universum, by telling us what makes the thing being that what it is (and not something else). Now, one could well accept this methodological regulative for *theoretical* terms and still deny its relevance for the term 'quale' because it is not a theoretical term, unlike "fraction" or "atom". After all, "quale" is to denote the very quality of our mental states which is the basis of all our judgements about identity (something like sense-data or raw intuitions), since the identity criteria of any theoretical object (except for the *ideal* theoretical objects like numbers)

[8] See [Dennett(1988)].

[9] They are intrinsic because they are real and irreducible to extrinsic properties of mental states; they are ineffable because otherwise the possibility of knowing what it is like to be a bat would be open (all it takes is teaching the bat, or another being using echolocation, how to describe it); they are apprehensible in consciousness because they are defined as "felt qualities"; and finally, they are private because they are supposed to be the residue of a mental state when it is stripped of every overt manifestation, they are the intrinsic qualities of *my* mental states.

[10] Ibid.

[11] Note that 'quale' is a philosophical term that belongs to metaphysics of mind; and metaphysics is a theory *sui generis*. The whole idea comes, of course, from Quine's "No entity without identity."

must be ultimately stated in terms of *observables*[12] that themselves have no deeper epistemic foundation (otherwise an infinite regress would occur). So, the argument would go, we must rely on our ability to discern qualia without knowing their identity criteria, for otherwise we could not even make sense of judging whether some observable conditions are fulfilled.[13]

The outcome of the previous hypothetical discussion is that the requirement of identity criteria for qualia may be too strong, for they may not be *merely* theoretical objects. They are theoretical insofar they are stipulated as a part of the *explanandum* in the mind-body problem, and they are not theoretical insofar they are experienced. Remember, however, that the problem addressed by the explanatory argument is theoretical, not experiential: I do feel something but I do not know how to isolate the felt quality itself; and that seems to be necessary if we are to try to explain it in scientific terms. Therefore identity criteria are still relevant. Imagine somebody wanted you to explain something you cannot recognize. Explain THIS! What? Well, THIS thing I am referring to. Ok, my explanation of THIS is ... (if THIS is what you referred to). No, you have explained THAT, but not THIS. How am I to know?[14]

Let me finally turn attention to showing why qualia lack identity criteria. Actually, all we need is to consider Wittgenstein's argument against private language. Given the salient features of qualia mentioned above, especially privateness and ineffability, it seems to be clear that the possibility of private language goes hand in hand with the possibility of identity criteria.

> §258. Let us imagine the following case. I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign "S" and write this sign in a calendar for every day on which I have the sensation. – I will remark first of all that a definition of the sign cannot be formulated. – But still I can give myself a kind of ostensive definition. – How? Can I point to the sensation? Not in the ordinary sense. But I speak, or write the sign down, and at the same time I concentrate my

[12] At least so in the empiricist paradigm which I have taken over from the whole discussion about qualia.

[13] This hypothetical counterargument is, I think, wrong, though it is far from obvious. It rests on a wrong assumption about cognition inherent to empiricism (the Myth of the Given) that was revealed by W. Sellars in his *Empiricism and the Philosophy of Mind*. Anyway, it is charitable enough to grant a fallacious argument to a hypothetical opponent.

[14] It might be rightly objected that theorists dealing with the qualia *do* understand what is to be explained, since even some neurophysiologists have recognized qualia as forming the hard problem. What they recognize, however, as the *desideratum* of their theories is not an account of particular qualia but of subjectivity in general.

> attention on the sensation - and so, as it were, point to it inwardly. – But what is this ceremony for? for that is all it seems to be! A definition surely serves to establish the meaning of a sign. – Well, that is done precisely by the concentration of my attention; for in this way I impress on myself the connexion between the sign and the sensation. – But "I impress it on myself" can only mean: this process brings it about that I remember the connexion *right* in the future. But in the present case I have no criterion of correctness. One would like to say: whatever is going to seem right to me is right. And that only means that here we can't talk about 'right'. [Wittgenstein(1953)]

One of the many things this paragraph shows is that the reality of how things seem to us is not independent of our judgements about how they seem. As Dennett likes to put it, there is no *real* seeming. Wittgenstein claims that saying "I know I am in pain." violates the grammar of the verb "to know" as ordinarily used, at least if I thereby mean something more than just "I am in pain." Thus the best sense Wittgenstein can give to the beforementioned expression is that it does not make sense to doubt the occurence of one's own feelings.[15] In the Wittgensteinian sense, where there is no possibility of a mistake, there can be no knowledge either. Consequently, we do not find out what we feel, we simply report what we feel. However subtle the difference may seem, its consequences are important. For as Wittgenstein says when he replies to a hypothetical accusation of behaviourism, our expressions of feelings do not denote special objects, but they *express* them.[16] It is helpful to recall the famous beetle-in-the-box analogy. If we maintained that verbal expressions of feelings do denote some things everyone is familiar with only from their own case (this satisfies the delineation of qualia), then

> [s]uppose everyone had a box with something in it: we call it a "beetle". No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But suppose the word "beetle" had a use in these people's language? – If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a *something*: for the box might even be

[15] Cf. [Wittgenstein(1953), §246].
[16] Cf. [Wittgenstein(1953), §244].

> empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
>
> That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant.[Wittgenstein(1953), §293]

The last remark reflects Wittgenstein's anti-essentialist approach to language. We should not expect that language works in the same fashion no matter what language-game we actually play. Thus the grammar of words like "table", "stone" may not be the same as that of "love", "pain" or "anxiety", even though they are all substantives and often appear in sentences with identical surface structure. Surely, Wittgenstein's account is susceptible to critique, especially when taken positively rather than negatively, but let me conclude, for the time being, that expressions of feelings do not necessarily acquire their meaning by naming something we are intimate with, i.e. qualia.

A defender of the concept of quale could now reply: allright, you've just brought Wittgenstein's argument to the conclusion that our language-game 'Feel it – Report it' can be played independently of what *it* is, but why could not words like "pain" or "itch" denote something once the usage of the words is settled by practice? That means: after a child learns to report "I am in pain." instead of crying, it nevertheless comes to know what pain is because the word helps it to hold the content (the *quasi*denoted quale) fixed. After all, it would not report pain consistently *if the reported thing were not constant.* Hence, the quale is that what induces the report. Well, is the defender better off? This argument suggests that identity criteria are stated in terms of identical reports, i.e. identical verbal expressions. Note the emphasized assumption in the last but one sentence: taking it as a matter of *fact* whether the reported thing is constant presupposes there are identity criteria. It means, effectively, that we need criteria of consistency in usage, other than just stipulating the reported thing constant (for in that case the reality of qualia would directly depend on our judgements, and hence qualia would not be *intrinsic* properties of our mental states). What could those consistency criteria be? Presumably, the consistency of the usage is judged by the language community that teaches it. And this is ultimately done on the basis of observable and hence non-private facts, viz. Wittgenstein's discussion of when we say that somebody *knows* how to continue an arithmetical series.

Wittgenstein notes, in the quoted paragraph, that the thing in the box, the beetle, could be even constantly changing or the box might be empty.

Were the box labelled "consciousness" instead of "beetle", this might express the familiar idea that at some level of our minds there is a united 'stream of conscious contents'. I think it is the unity of the stream (consciousness) which underlies most of the appeals to and intuitions about some quality of experience that is hard to explain. But the unity is a quality of consciousness as a whole, whereas qualia are qualities of mental states, i.e. of differentiated contents of consciousness. The previous discussion has hopefully shown that the type-identity of linguistic expressions does not ensure the identity of intrinsic quality which is supposed to be part and parcel of what induces the subject to report this rather than that. Furthermore, if mental states, qua Kantian judgements, are differentiated by their conceptual structure, we still cannot be sure that the quality that accompanies each type of mental state is the same, unless it is stipulated that identical conceptualization occurs only if the quality of experience is identical. And if a defender of qualia stipulated the quale to be identical for every type of mental state then qualia would not be inexplicable by structure and function only, since the proponents of qualia would admit, as far as I know, that the content of a mental state (which governs its identity and under the stipulation also the identity of qualia) can be functionally analyzed.

The last option for a defender of qualia seems to be appealing to a sort of intuition. Do we not know what it is like to see red (unlike green, for example), and could we not find out that the quale is identical simply by concentration? Not quite, for concentrating one's attention does not work as ostension, which is mentioned in the already quoted §258.[17] Anyway, the defender may still claim she 'simply knows' what quale she is experiencing, but since it is by definition ineffable, she cannot describe it in more details. However, there are *facts* about qualia and some of these facts can be articulated.[18] If we had intuitive access to the identity of felt qualities, then there should be no controversy about such simple facts as "These two qualia are identical." Such a controversy occurs, however, in respect to the change-blindness phenomenon which occurs when a subject of an experiment is watching an image a part of which is changing, but the subject is not aware of any change. The change can be either gradual (e.g. the background color slowly changes to another) or abrupt, with a short mask between the two images. Since the subject is not aware of the change, she believes she is in the same mental state as before. The question that invites controversial answers is whether

[17] And even if it worked as external ostension, the general problem of ostensive definition would arise: in order to recognize what is pointed at, we would already have to know the grammar of the introduced term. And if we were to know it, would the definition get any further than the stipulation "qualia are qualities of mental states"?

[18] Cf. [Nagel(1974), pp.393,396].

the qualia are identical before and after the change, or whether they are different, even though the subject does not notice. The latter option seems to be untenable from the start, since if qualia could change without a notice, we have no special epistemic access to them, which contradicts the original assumption of proponents of qualia. Regarding the first option, it is hard to find other reason for holding the qualia identical than that the subject *thinks* the images are identical. Either way, the very fact that defenders of qualia are not unanimous about the answer suggests the concept of quale is too vague to be allowed as a part of the *explanandum*.

Taking the argument against qualia a bit further (to a final stand, I daresay), we can claim that qualia, as they are intended by their proponents, cannot be empirically real in the Kantian sense. Qualia are said to be ineffable qualities of experience, hence they are not conceptualized. As such, qualia cannot be known by us, for empirical knowledge is a synthesis of sensibility and understanding which imposes concepts on intuitions of the former.[19] I take qualia in their intended meaning to be, in Kantian terms, very close to pure intuitions. Yet every conscious being allegedly *knows* what it is like to be it. Are we supposed to allow for a special kind of non-conceptual knowledge (if we wanted to pursue a charitable interpretation of qualia)? We can readily dismiss knowledge-how as a proper candidate, since there is no practical employment of that knowledge to think of. So, perhaps knowledge by acquaintance? "I know what it feels like being in pain, because I felt pain." However, this is not an expression of knowledge proper but rather a claim of the ability to categorize various sensations under different concepts. Knowledge by acquaintance is thus not knowledge of some quality but recognition of something as the thing it is (and not something else) *and*, by the same token, the ability to recognize it as identical in future. Being acquainted with a feeling does not amount to knowing its quality (which is, probably, supposed to be the *matter* thanks to which the feeling is recognized as such and such), but to be able to discern it as such among other feelings. Fortunately (and necessarily), we are *able to* discern things even if we cannot articulate on what ground we do so. This happens in the process in which intuitions are brought under concepts. If we could specify the ground of the recognition, it would mean we have already discerned *parts* that the original

[19] What if the proponents of qualia argued that knowledge of what it is like to be oneself is not supposed to be empirical but rather *a priori*? We could answer that Kant gave an account of what it is like to be a man when he stated the necessary features and conditions of experience. But that is not ineffable at all. And it would be wrong to retaliate that perhaps some bat-like Kant could equally specify the character of bat experience, for Kant's account was not limited to experience of man as a species but to experience in general (though as *intelligible* to *us*).

thing depends on.[20] So either we know some qualities of certain feelings, and then the qualities are intersubjectively understandable, since they are necessarily conceptualized, or we do not *know*, in the proper sense, the quality of a feeling, we simply feel it.[21] Once again we come to the conclusion that qualia cannot be known because they cannot be conceived of – only 'felt'. This time, however, we may exclude them (though only in Kantian framework) from the domain of empirically real things, since they can never be objects of experience.

At the end of this section, I would like to emphasize that the previous disqualification of qualia (as some intrinsic qualities of experience) does not mean we do not feel anything or that feelings are completely reducible to behaviour. It only means that reification of the non-conceptual character of experience is utterly inappropriate and that the concept of quale is therefore an idle wheel in the philosophy of mind and in any empirical theory whatsoever. Wittgenstein comments on it, rather cryptically though, as follows:

> §304 "But you will surely admit that there is a difference between pain-behaviour accompanied by pain and pain-behaviour without any pain?" – Admit it? What greater difference could there be? – "And yet you again and again reach the conclusion that the sensation itself is a *nothing*." – Not at all. It is not a *something*, but not a *nothing* either! The conclusion was only that a nothing would serve just as well as a something about which nothing could be said. We have only rejected the grammar which tries to force itself on us here. [Wittgenstein(1953)]

Qualia cannot be accounted for not only by science but by any other means pretending to aim at knowledge.

2.2 Transcendental self-consciousness

Coming back to our starting point, namely the explanatory argument, is there any other candidate feature of consciousness that would justify the second premise? There is a feature that many appeal to in an attempt at an ultimate argument for the irreducibility of consciousness: all experience is united in a single point of view that is by the same token aware of itself.

[20] The arguments is essentially the same as Sellars argument against the Myth of the Given.

[21] Throughout the argument I used a specific feeling of pain instead of general what-it-is-likeness of being a man which is employed in Chalmer's definition of consciousness. I did so mainly for clearer illustration and I am convinced that if the argument holds for a specific feeling, it *a fortiori* holds for the general what-it-is-likeness.

Let us first remind ourselves of the intuition supporting the idea of self-consciousness as an irreducible feature. C. McGinn, for example, wonders: "How could the aggregation of millions of individually insentient neurons generate subjective awareness?"[22] In the rest of the article McGinn argues that even though brain *is* the causal basis of consciousness, we may never understand the link because it is *cognitively closed* to us. Indeed, there seems to be an insurmountable conceptual gap between neuroscientific account of brain states and phenomenological account of mental states and subjectivity in general. The crux of the divide seems to be perspectiveness – while empirical sciences investigate their objects within causal order, looking for general laws governing transition from causes to effects, phenomenology[23] conceives of its objects as necessarily linked to a particular perspective or point of view. How could perspectiveless physical parts form a perspectival whole uniting *different* mental states? This appears to be beyond our understanding.

If one is convinced of causal dependence of consciousness on brain states, it is natural to look for a place in the brain that provides for this perspectiveness – a kind of bottleneck through which every afferent signal must pass in order for one to be conscious of it, a Cartesian pineal gland, the embodiment of the self. Natural and tempting as it may be, it is pointless, as D. C. Dennett convincingly shows in his *Consciousness Explained* and elsewhere. Not only is it logically fallacious, he argues, to assume there is a Cartesian theatre in which the self is watching presented percepts and thereby becomes conscious of them, but it can also be empirically refuted that there be a specific part of the brain responsible for turning unconscious information into conscious one.[24]

2.2.1 Kantian variations on consciousness

Kant provided us with a particularly suitable account of the problematic feature of consciousness, usually referred to as "transcendental unity of apperception". It is suitable for it allows to clearly distinguish between what we can call empirical and transcendental self-consciousness. Empirical self-consciousness consists of inner experience of oneself as an appearing object, i.e. it arises upon conceptualizing intuitions of one's inner states. Such self-consciousness is no knowledge of a thing in itself, it is knowledge of how we appear to ourselves, to phrase it in Kantian terms. The related concept of an empirical self has much in common, I think, with Hume's bundle self, for Kant remarks that empirical consciousness is "by itself dispersed and with-

[22][McGinn(1989)].
[23]I thereby mean any reflective account of mental states or consciousness in general.
[24]For the detailed argument see [Dennett(1991)].

out relation to the identity of the subject."[25] Concerning the mind-body problem, the empirical self-consciousness seems to be no greater mystery than consciousness of any outer object. Intentional structure of both inner and outer experience is the same, so the sole difference between them is the special status of the object to which properties and features perceived in inner experience are ascribed, namely the (empirical) self or person (with its history, character etc.). If we grant that AI researchers can, in the relevant sense, make machines (such as Mars probes) *aware* of their surroundings in which they can indentify distinct objects, we should easily concede the possibility of awareness of their internal states, which they do in fact keep track of for self-preserving purposes.

Transcendental self-consciousness, on the other hand, is of entirely different nature. The meaning of the term "transcendental unity of apperception" may be elucidated by Kant's well-known expression that "it must be possible for the "I think" to accompany all my representations, if they are to be anything to me."[26] That is, the united manifold of representations presupposes transcendental self-consciousness that implicitly relates all representations to one subject.[27]

The attribute 'transcendental' signifies that the term refers to conditions for the possibility of experience in general and of empirical self-consciousness in particular. Kant explains it as follows:

> The synthetic unity of consciousness is therefore an objective condition of all cognition, not merely something I myself need in order to cognize an object but rather something under which every intuition must stand *in order to become an object for me*, since in any other way, and without this synthesis, the manifold would not be united in one consciousness. [Kant(1781), B 138]

P. F. Strawson elaborates the idea behind the term:

> [I]f different experiences are to belong to a single consciousness, there must be the possibility of *self*-consciousness on the part of the subject of those experiences. It must be one and the same understanding which is busy at its conceptualizing work on all the intuitions belonging to a single consciousness, and it must be possible for this identity to be *known* to the subject of these experiences.[28]

[25] [Kant(1781), B 133]
[26] [Strawson(1966), p. 93].
[27] Cf. [Kant(1781), B 132].
[28] [Strawson(1966), p. 93].

Strawson then traces and supplements Kant's findings on the transcendental self-consciousness in a way that yields many conclusions worthy of our attention. Since my aim here is not to present an accurate account of Kant's view of the matter, but rather to isolate the sense of consciousness relevant for our discussion of the Hard-problem, I will only make a list of the characterics which give a clearer view on the meaning of the technical term.

1. If it is to be possible for the "I think" to accompany all representations, it is necessary that the order and connectedness of experiences is recognized as different from the order of the objective[29] world. That means that only thanks to the distinction of the subjective and the objective order can I recognize myself as one possible experiential route among others, one track of point of view of the objective world. The distinction itself is possible only if one can get something wrong in perception of the objective world. Otherwise it might never occur to me that I am not a god-like consciousness affected by other things than my own states.[30]

2. "In the synthetic original unity of apperception I am conscious of myself not as I appear to myself, nor as I am in myself, but only that I am. This representation is a thought, not an intuition."[31] Therefore, the transcendental self-consciousness cannot be an object of experience. Furthermore, the content of such a thought of self-consciousness is general: it is about consciousness as such, not about the particular (empirical) consciousness individuated by particular experiences.

3. The "I" in the "I think" of the transcendental unity of apperception to which experiences are ascribed bears no criteria of identity, as it makes no sense to doubt to whom this or that experience occurs. However, the "I" acquires specific referent thanks to one's body, which is an object of outer sense, being the means of identification of selves in interpersonal communication. Thus it is because I am involved in communication with other people, to whom I wish to convey my *stance* and in whose eyes I am identified with specific object of outer sense (my body), that my representation of the "I" has a specific referent instead of a general concept of the subject.[32]

[29] "Objective" in Kantian sense, i.e. as an order of appearances of objects of outer experience, not an order of things in themselves.

[30] Further discussion of the issue can be found in [Strawson(1966), pp. 97-112].

[31] [Kant(1781), B 157], quotation owing to [Strawson(1966)].

[32] It could be argued that perhaps *other* people identify me with my body just by

Strawson comes to the conclusion that Kant's explanation of the possibility of transcendental self-consciousness is based on the claim that we are conscious of our synthesizing activity (whereby intuitions are brought under concepts) or at least of our power to do so.[33] "[O]ur consciousness of the identity of ourselves is fundamentally nothing but our consciousness of this power of synthesis, ..., and of its exercise."[34] It will be argued in section 5 that "the consciousness of this power of synthesis" arises specifically upon semantic considerations about judgements or claims that one is about to endorse.

2.2.2 Empirical consciousness

The importance of the distinction between empirical and transcendental self-consciousness lies in the fact that the latter is no object of experience: it is a necessary feature of the phenomenon of consciousness, not some pure mind of its own. Therefore, in trying to explain scientifically the mystery of consciousness, we should not be led astray and look for an empirical account of an *entity* called consciousness (the purely thinking "I"), because there is no such entity. However, the features which Kant specifies under the term 'transcendental' may have natural foundation which we may be cognitively open to.[35] What I am suggesting is this: consciousness, as an *explanandum* of natural sciences, is not an entity with powers too mysterious[36] to be accounted for by current scientific concepts; it is rather a feature distinctive for human mind, thattself is best understood as a file of cognitive faculties

extrapolating the identification from their own case, i.e. that the identification with one's body is primary and identification of the others is only derived. I doubt it could be so. We only need to realize that every communication is somehow embodied. If we were telepaths with no need for overt communication, we would probably find it a mere stipulation that selves should be identified with bodies.

[33]Kant's version of the same claim can be found at B 133.

[34][Strawson(1966), p. 94].

[35]Note the change in force of the intuition that drives McGinn to his conclusion about cognitive closure: once the hard-to-explain matter is a feature (and not an entity), it seems more likely to be accountable for by science, since the feature specifies a conceivable function and we know that the same function may be realized by various means. That is, once we concede that we need to know how the "I think" can potentially accompany every mental state, we know at least what we are looking for and may even have a rough idea how to get to it thanks to our knowledge of similarly structured problems (how can "... is true" accompany any declarative sentence? - by stating that S and "S is true." are equivalent in meaning).

[36]The prime example of such power, that is regarded as beyond contemporary scientific conceptual scheme, is intentionality. This stance together with the conviction that consciousness is a natural phenomenon leads to attempts (for example, by John Searle) to establish intentionality as a new non-analyzable physical concept beside force, charge etc.

which we can recognize and which are exhibited by a perceivable being. And unless we find reasons to think that the faculties may be exhibited by, for example, organisms only, we should not restrict the domain of things eligible for consciousness.

In my view then, consciousness is a term much more specific than mind: it refers to our[37] potentiality of elevating any mental state to a higher level of attention by realizing that the mental state occurs to oneself. Thus I propose to interpret Kantian phrase about "I think" possibly accompanying all my states as suggesting a higher-order theory of consciousness according to which to be conscious of X is essentially to entertain a higher-order thought that I think/perceive/believe X. Such an interpretation is at most a first approximation rather than a satisfying explanation, for many important issues are left unresolved. Does a higher-order mental state need to actually occur in order for us to be conscious of its lower-order content? In what medium of representation is the higher-order thought *about* its content? In virtue of what do we possess the capacity for higher-order thoughts? Or is it an unanalyzable cognitive function? How is consciousness *of mental states* related to *transcendental* consciousness? P. Carruthers puts forward a theory that tries to answer some of these questions. For an outline of the theory and to get a better idea of what higher-order theories of consciousness involve, see appendix A.

A simpler but circular explanation would be that consciousness refers to our potentiality to bring any mental state before the mind's eye, to get conscious of it. It is a peculiar feature of the concept of consciousness that any attempt at its straightforward specification defies non-circular definition (or definition in which the explanation is postponed to the definition of another obscure term, like "attention").

2.2.3 Natural foundation of transcendental features

It is necessary now to shed more light on the relation between the transcendental character of the concept of consciousness as I intend it to employ it in this work and its purported empirical foundation. To repeat: the transcendental unity of apperception represents a necessary condition for there to be a mind having experience as Kant understood it. In the minimal interpretation, the property that it must be possible for the "I think" to accompany any mental state is a *logical* property of mind, rather than empirical. Thus

[37] While referring explicitly to human minds now, I do not preclude other beings' minds to happen to exhibit this potentiality as well. The specific reference is only to help the reader understand the content of the term, for it is generally held to be a distinctive feature of humans among various beings nowadays considered as having a kind of mind.

it does not follow that *consciousness of* something is caused by occurence of a mental state of the form "I think of x." But if it is not an empirical property, what sense does it have to look for its natural foundation?

Here we have to plunge deeper into Kant's theory apperception. According to Kant, we are transcendetally conscious of ourselves by the very act of representing, not by representing ourselves as an object of inner sense (which gives rise to empirical self-consciousness). Again, the "I" in this self-consciousness is only apperceived, and thus it is devoid of any properties which are otherwise, i.e. in ordinary perception (both inner and outer), means of identification. Meaning of the "I" of apperception is that it is *the* logical subject in which all acts of synthesis are united, which is the condition of possibility for unified experience. Kant stresses the merely logical character of unity:

> The identity of the consciousness of myself at different times is ... only a formal condition of my thought and their coherence, and in no way proves the numerical identify of my subject. [Kant(1781), A 363], cited by [Sellars(1974)]

As Sellars subsequently points out, this "suggests the possibility that successive acts of thought might belong together, as acts of the same I, and yet be successive states of different noumenal subjects."[38] For us, the important consequence is this: somehow the acts of synthesis are *necessarily* united in the transcendental "I" by having the property (which is not stipulated but transcendentally inferred) that one is conscious of them by doing them. The (inter)dependence of the "I" and acts of synthesis is logical, not causal. Finally, the answer to the question about natural foundation of transcendental features is following: the acts of synthesis are open to empirical scrutiny because judgements (the outcome of synthesis) can be expressed in language. Therefore, it might be meaningful to investigate whether taking acts of synthesis literally as utterances could in any way make the transcendental relation of consciousness and synthesis clearer. Certainly, this holds only if we accept that the conceptual is tightly linked to the linguistic.

2.3 The conceptual and the linguistic

Intuitions are brought under concepts in the act of synthesis. The standard account of Kant's Copernican revolution is that things appear to us as they do because our cognitive capacities inevitably impose certain coherent conceptual framework on things in themselves. Some of the conceptual

[38] [Sellars(1974)].

structuring is intrinsic to human mind, namely that which corresponds to categories as arrived at by the Metaphysical deduction.[39] The rest, it can be said, is derived or acquired. Now, I am convinced that the sense of the term "concepts" is best explicated as an abstraction of *meanings* of type-identical linguistic expressions from their symbolic embodiment[40] and the circumstances in which they were used. By illustration, we often express meaning of a linguistic expression by means of concepts: "The word 'le chien' means DOG." This is truly understood as an explanation, not just as a rule of translation "le chien" → "the dog". Thus I am commited to saying that whoever has mastered language is also capable of conceptual structuring. Or to state it as a logical relation: the linguistic is *sufficient* for the conceptual.

This may seem very bold initially, but it is rather a trivial statement under the following intepretation. Mastering language requires that the speaker recognizes rules that govern the use of linguistic expressions (which does not imply that she always abides these rules). A law-like disposition of a digital thermometer to produce certain strings of symbols, or conditioned articulated sounds of a parrot, are not linguistic expressions proper. The rules themselves are all that is needed to have a working concept. So the linguistic is sufficient for the conceptual because in order to be a speaker of language, one has to internalize rules of use - and these are constitutive of concepts.

The reverse is true as well: the conceptual is *sufficient* for the linguistic. If one is a being among other beings in the world, all it takes to have a language is to assign each concept some unique symbolic representation such that it can be easily recognized by one's peers' perceptual faculties. That is, it only needs that vehicles be construed that would convey the concepts.

It follows from mutual sufficiency that the linguistic is necessary for the conceptual and vice versa, and therefore they are (logically) equivalent. In effect, the link is now as tight as it can be - perhaps so tight that one doubts whether the distinction between the conceptual and the linguistic is still worth preserving. I propose the reader accepts it now as a working hypothesis with the prospect that it will be argued for in section 3.1, along with its application to the explanation of the relation between language and consciousness.

---

[39] Its intrinsic nature follows from the inherence of our cognitive capacities: "Kant's revolutionary move was to see the categories as concepts of functional roles in mental activity. Categorial concepts are not, indeed, innate. They *are* formed by abstraction, *not*, however, by reflecting on the self as object, but by reflecting on its conceptual activities." [Sellars(1974), p. 68].

[40] I thereby mean both sounds and spatially extended signs.

2.4 The Hard problem revisited

Let me return to our original point of departure, the Hard problem. I have argued that identifying qualia (or what-it-is-likeness) as the really hard part of the mind-body problem is conceptually wrong because we do not even know what exactly they are, not to mention what role they play in our mindedness; thus I rejected that the concept 'quale' could substantiate the second premise of the explanatory argument. I have brought together some findings on Kant's concept of transcendental self-consciousness in an attempt to 1) isolate the aspect of the mind which makes the mind-body problem hard and confusing; and 2) make clearer what actually needs to be explained, what the mystery of consciousness is about. Finally, in the previous section I indicated that the process of synthesis could be studied at the explicit level of linguistic judgements. So, recognizing transcendental self-consciousness as the hard part that we should focus on leads us to study how concepts are employed; because, to repeat once again, the transcendental self-consciousness is constituted by the very doing of synthesis. And we can study it through studying language, under the simplifying but, as I hope, fruitful assumption of the tight link between the linguistic and the conceptual.

The above outlined reformulation of the Hard problem is safe, I believe, from Chalmers' argument aimed against the theorists inclining to functionalism who allegedly tend to reduce the Hard problem to a set of easy problems of how our cognitive abilities (functionally specified) are realized. While it is true that Kantian transcendental self-consciousness is described in terms of structure and function (which is a necessary consequence of *transcendental* inference), explanation of its natural foundation seems to be no easier than explaining the original what-it-is-likeness of experience – it is only more specific and therefore intelligible, which is surely a merit, not a drawback.

What makes the newly stated Hard problem really hard now is *meaning*. It is generally acknowledged that human sapience consists in operation on meanings, be they recognized in words or acts. Despite many efforts, we still lack a satisfying account of meaning. We conceive of meaning as of that which governs the use (and even misuse) of linguistic expressions in order to convey a message. To the best of my knowledge, there is no theory that would convincingly explain the constitution of meaning or its nature. On the other hand, substantial facts about language and communication have been revealed during the pursuit of meaning's nature which can be found relevant for consciousness. On the following pages, I will try to assemble the relevant findings while keeping them nested in the theories they come from so that the context, in which they were conceived, ensures that their relevance will not be exaggerated.

3 Concepts and compositionality

The conceptual exhibits (being tightly linked to the linguistic, as discussed above) various features some of which may be highly relevant for transcendental self-consciousness. There are two major issues in relation to which I will discuss the features: compositionality and the nature of meaning. Before I tackle these issues, however, let me briefly specify some ontological constraints of the conceptual as I understand it.

3.1 Some ontological remarks about the conceptual

Experience arises when intuitions are brought under concepts. Experience is undoubtedly individual, and so are intuitions, being yielded by sensation. What about concepts? Both options seem to be possible. Either we can say that everybody employs her own concepts, applying, as it were, individual conceptual framework on the undifferentiated matter of sensibility. Or we can take conceptualization to be a relation of a subject to her individual intuitions and *universal* concepts. Both options face their difficulties. If concepts are taken to be individual (for example, as mental particulars), it is difficult to account for intersubjective understanding: do we not mean the same thing by saying/thinking that the cat is on the mat? There is no way to establish whether someone's judgement that the cat is on the mat is *in fact* someone else's judgement that the undetached cat-part is on the mat.[41] For there is no metalanguage (independent, universal) of thought into which different conceptual frameworks could be reliably translated and thereby compared with each other. On the other hand, if concepts are held to be universal, such that individuals (precisely: their faculty of understanding) relate their intuitions to them, it is difficult to account for the basis of disagreement. If I claim "This shirt is blue." while someone else claims "This shirt is grey.", is it because we differ in intuitions, or in concepts tokened by our faculty of understanding, or in our word-concept associations? Any answer to this question would be arbitrary. To prefer one choice over the others is to prefer one locus of semantic relativity over others: the cause of disagreement may be identified as a difference between men's sensibility, understanding or linguistic capacity. Surely, arguments can be presented in favour of any locus, so that the choice does not appear arbitrary, but every argument, I daresay, would be inference to the best explanation, where "the best" is again

[41] The allusion to the radical indeterminacy of translation actually does not show that there is some inherent inconsistency in the conceptual thus conceived. It rather shows that the conceptual is of no or little use in explaining meaning and mutual understanding, for all that is *comparable* are linguistic expressions.

to be judged according to one's preference about what should be considered *constant* among people.

If we take seriously the previous paragraph, together with the proposed tight link between the conceptual and the linguistic, why do we not abandon talking about concepts at all? What explanatory function does the conceptual serve that we cannot dispose of it in favour of the linguistic only? Intuition suggests that perhaps not all experience is articulated (mentally) in words and yet it is logically structured in such a way that it can serve as a premise in inference, or that it can eventually be articulated if the right words are chosen. This suggestion reveals commitment to the conceptual as the language of thought which is usually accompanied by the claim of semantic primacy of the conceptual (at least in the order of explanation), i.e. meaning is primarily property of thoughts and words acquire it only derivatively by expressing them. However, the language of thought hypothesis is an extraordinary semantic theory and concepts can be dissociated from words and meanings even if we do not adopt it. In different perspective, concepts can be identified with meanings: concepts are abstract entities with properties that determine the correct use of words that express them. The conceptual so conceived is employed in purely semantic theories which are not concerned with psychological aspect of language production; they account for the meaning of utterances, not for the process the outcome of which is the utterance itself. In the Kantian paradigm adopted here, however, we need to account for the *individual* act of synthesis. So even if we take the second option, according to which concepts are universal, that is in accord with the meaning-concept identification, there are still individual tokenings of concepts that do not occur haphazardly but according to rules (embedded in understanding). There is lot of empirical evidence in support of the fact that people differ in those rules. It is most clear in cases of classification. I can conceptualize a blindworm as SNAKE while someone else as LIZARD; and it does not necessarily follow we have different opinions on what should count as a snake or a lizard - I can be eventually corrected that the blindworm is actually a lizard if I conform to the same regulative force of the meaning of "snake" as the other. Still, my idiosyncratic rules may classify the percept as of a snake.[42] Therefore concepts may represent universal meanings but there

[42]What if somebody argued that my conceptual classification as either a snake or a lizard is indeterminate until I make the judgement, i.e. articulate it explicitly as "This is a snake." (be it overtly or in one's mind)? For then I could either apply a word correctly (i.e. in accord with its meaning) or incorrectly, because I do not know its meaning (and thus I could not be persuaded logically that it is a lizard - only one's authority could make me concede it). But this would preclude the possibility of self-correction which often happens. Also, the conceptual classification may clearly occur before anything that

still have to be individual rules of concept application. The qualification "individual" means that the rules must be realized in every individual, not that they must differ. In the naturalist approach to the mind, this amounts to saying that the rules, despite their functional generality, are implemented by cognitive mechanisms specific to each individual. But even under the functional specification the rules may differ, as the idiosyncratic aspect of language use suggests. This is the reason to keep "the conceptual" in our explanatory vocabulary, distinct from words and meanings which *must* be thought of as universal (or common).

3.2 Compositionality, productivity, systematicity

The capacity to understand and produce sentences which have never been encountered before is a distinctive feature of human mind, usually referred to as 'productivity'. Productivity represents a strong reason to accept that meaning in natural languages obeys the principle of compositionality: the meaning of a complex expression is determined by its structure and the meanings of its constituents.[43] Thus productivity is achieved even with finite sets of semantically primitive expressions and grammatical (syntactical) rules. Finally, systematicity is a feature of language that also supports the principle of compositionality: there are discernible patterns among sentences such that apprehending a sentence under some such pattern is often sufficient for understanding a sentence of the same pattern with different constituents with which the subject is familiar.[44]

3.2.1 Compositionality - a fact or a principle?

Neither productivity nor systematicity are conclusive for compositionality - first, it is (only) inference to the best explanation that yields compositionality out of either of the two, and second, there are notorious counterexamples from

we could call articulation: my sudden alertness is due to the belief that snakes can be poisonous (while lizards are less likely to be) and the belief that what I have encountered is snake.

[43]Definition taken from [Szabó(2004)].

[44]Clearly, this holds only if the pattern of a sentence is recognized as well. We are not concerned here with the process of language acquisition, at some stages of which it may happen that a child *understands*, say, "The cat is on the mat." and does not understand "The dog is on the table." even though it knows dogs and tables (now entertaining 'understanding' in a rather relaxed sense). Probably there must be first some conception of the meaning of a sentence as a whole than the general pattern can be recognized through acquaintance with other examples - especially if sentences are the basic units of semantic evaluation.

natural languages that seem to violate the principle. A brief discussion of one type of counterexamples (which attracts perhaps the most of attention) will show that the status of compositionality depends, unsurprisingly, on the conception of meaning. Consider two sentences:

(a) Carla believes that eye doctors are rich.

(b) Carla believes that ophthalmologists are rich.[45]

While one sentence can be true, the other may be false, despite the fact that the embedded clauses are semantically equivalent because 'eye doctors' and 'ophthalmologists' are synonyms. This example violates compositionality - but only in the framework of *truth-conditional* semantics, because meanings of the complex sentences (their truth value) do not compose in the same way even though meanings of their constituents are identical. Should we conclude that compositionality does not hold or that truth-conditional semantics is inadequate for natural languages? Or that special semantics has to be designed for propositional attitudes whose context is referentially opaque? The fact that semanticists will rather try to amend their theories than give up on compositionality suggests that it is a feature worth preserving in any theory of meaning.

But the issue may be taken even further: does compositionality follow from the nature of meaning, or is it a regulative principle that determines meanings? To appreciate the question, consider that compositionality as defined above is compatible with an almost reverse claim, called Frege's *context principle*: the meaning of an expression is determined by the meanings of all complex expressions in which it occurs as a constituent.[46] Thus compositionality seems to fit both atomistic and holistic views on semantic primacy (i.e. words or sentences); it is only if compositionality is interpreted as expressing the causal or explanatory order in the domain of meaning that semantic atomism can be deduced from it.[47] If meaning of an expression is a matter of fact *independent of the principle of compositionality* (like in purely referential semantics) then compositionality may be *demonstrated* to be true. For it may turn out that compositionality either holds or not. Clearly, any violation of compositionality may be due to a wrong or incomplete model of meaning,

[45]Examples taken over from [Szabó(2004)].

[46]Ibid.

[47]Szabó points out in his (2004) that 'determination' of meaning in the original definition (p. 30) is to be taken abstractly, disregarding its causal or explanatory connotations. The definition then says no more than that there is a function from meanings of parts and their structure to the meaning of the whole. A stronger, atomistic version of compositionality reads: complex expressions *have* their meanings *in virtue of* their structure and the meanings of their constituents. This is a factual claim.

but it should still follow, from the factualist stance to meaning, that compositionality is also a matter of fact. If, on the other hand, meaning is a matter of convention then it is the very principle of compositionality which enables us to abstract meaning of constituents of complex expressions by directing us to look for identical semantical contribution of the constituents everywhere they appear.[48] This too rough a dichotomy suggests that while the former, factualist consequence is congenial with semantic atomism, where emphasis is put on representation, the latter is congenial with semantic pragmatism and holism.

I owe the insight that compositionality might by understood as a *regulative* principle to J. Peregrin, who argues in his (2005) that the principle of compositionality is "(co-)constitutive of the concept of meaning, and thereby of the concept of language."[49] He characterizes the issue about the factual nature of compositionality as follows:

> [*E*]*ither* we can satisfactorily explicate the concept of meaning without the help of [the principle of compositionality], and then we are vindicated in taking the principle as an empirical thesis, *or* we cannot do this, and then the compositionality of meaning, and hence of language, is a con ceptual, 'apriori' matter. [Peregrin(2005), section 1]

Exploring Frege's work with its stress on semantic primacy of sentences over words and the central role of *truth* in semantic considerations leads to the suggestion that compositionality, in cooperation with other principles, *constrains* what meanings of subsentential components could be, i.e. what is their semantic contribution to the basic semantic value - truth of a sentence. The important outcome is that while we can stipulate what meaning is in whatever way we want (and then see whether compositionality obtains or

[48]The relation between conventionality and the regulative status of compositionality deserves perhaps a further explanation. If meaning is conventional then so is the relation of synonymy. If there are, as a matter of fact, no synonyms (which seems to follow from the conventionality of meaning, as Quine's discussion of the two dogmas of empiricism shows) and if every word-type can be assigned different meanings anytime it compositionally misbehaves, arguing that it is a case of homonymy, then compositionality is vacuously true according to the factualist reading (for every word-token can be assigned a unique meaning, and so it would be trivial to construct a function satisfying the compositionality constraint). So it would not make sense to understand compositionality as a fact that depends on some primary facts about meaning. The situation is reversed: meaning of an expression is established so that it conforms to the principle of compositionality.

[49][Peregrin(2005), section 1].

not), if we heed to the sense of "meaning"[50] then compositionality is already entailed.

3.2.2 Locus of compositionality

A question may be asked to what extent is compositionality confined to the linguistic only. Cannot we, after all, regard compositionality as a principle of organization intrinsic to the faculty of understanding? Could we not hold that part of our linguistic capacity, embedded in our cognition, is the ability to derive semantic contributions of subsentential expressions so that we can, in turn, employ them productively in forming new sentences? Let me spare few arguments in favour of a positive answer to these questions.

The whole matter (and so the very soundness of the question) rests heavily on the way we conceive of the relation between the linguistic and the conceptual. Since the principle of compositionality speaks explicitly about meanings of *linguistic* expressions, the principle can apply to (or hold in) something non-linguistic only in a metaphorical sense. To say, for example, that compositionality applies to the conceptual as well would be misleading, for concepts are the very range of semantic evaluation function.[51] We can hardly speak about meaning of a concept if concepts are to be, besides the individual aspect I exposed in 3.1, pure meanings. To be clear, it does not follow that all concepts have to be simple. The belief in the distinction between simple and complex concepts does not commit us to the belief that simple concepts *compose* into complex ones. Note that the stronger version of the compositionality principle states that meaning of the complex expression is a function of meaning of its constituents *and* their structure – but where is the structure of the conceptual to be observed? Perhaps all we can say about a complex concept, say PRESENT-KING-OF-FRANCE, is that it *consists of* concepts like PRESENT, KING and OF-FRANCE. The structure remains unknown, unless we *stipulate* that the structure copies that of the original linguistic expression.[52] But to suppose that would be to treat the

[50]Which definitely is not the range of an abstract semantic mapping function that can in principle contain any kind of objects.

[51]To my surprise, talking about compositionality of concepts is not rare. Fodor in his (2008) seems to take compositionality as originally applying to concepts, which are types of tokens in LOT. Even his critics (in [Prinz and Clark(2004)]) take on compositionality of concepts. If the only way to find out how thoughts are conceptually structured is to express them in a sentence, I don't see how compositionality of concepts could be non-derivative from that of words. Clearly, I am missing something.

[52]The problem does not disappear, however. It is only postponed to the discussion about criteria of primitive concepts. Is BACHELOR a simple or complex concept? If it is simple, what is the meaning of "A bachelor is an unmarried man." (how could the

conceptual as the language of thought. Indeed, then it could make sense that concepts compose, although it would still be a vacuous claim, for the language of thought *must* be compositional in order to account for productivity. To sum up, I think compositionality cannot be transferred from its natural habitat (the linguistic) without adaptation.

The adaptation which I propose consists in conceiving the principle of compositionality as a description of a necessary feature of our cognition. In the non-factualist reading, compositionality can be seen as a regulative principle according to which users of language abstract meaning of semantically incomplete expressions.[53] The regulative principle has to be hardwired in our cognition, for it must guide us in the process of language acquisition. One cannot be told that language is compositional before one masters it. And one could hardly realize it just by studying patterns of sounds or symbols unless one already had a propensity to conceive of meaning as compositional. Such propensity, together with the capacity to actually recognize semantic contributions of parts to the whole, would then be one of the prerequisites for language acquisition that Chomsky held to be innate.

3.2.3 Compositionality and reflexion

Finally, what is the purpose of adaptation of compositionality to the level of understanding? I suggested to take quite literally Kant's claim that it must be possible for the "I think" to accompany all my representations. Specifically, the capacity to prefix any representation with the "I think" is constitutive of consciousness. The actual operation of adding the "I think" is usually called 'reflexion', and it is a philosophical standard to claim that we can, in principle, reflect on any mental state, even on a reflective act itself.[54] It seems to me that philosophical tradition has often treated reflexion as so a fundamental capacity of human mind that it could hardly be explained in terms of something else.

Now, there is a feature of reflexion which invites an explanation based on linguistic considerations: every reflective act is unique, because the objects of reflexion are unique, and yet we claim they are yielded by a single capacity.[55]

definiens of a simple concept be complex?). If it is complex, can the conceptual structure be independent of the linguistic definition?

[53] I.e. expressions that by themselves cannot be assigned primitive meaning, e.g. subsentential expressions according to semantic holism.

[54] I do not thereby wish to commit myself to any particular ontology of the mind. You can read 'mental state' as anything you think that one may be conscious of - thoughts, representations, proposition, feelings etc.

[55] Compare with the capacity of driving: we tend to say that acts of driving cars and bicycles are exercises of *different* capacities.

This suggests that reflexion is a formal operation – its sense is independent of the meaning of the 'mental content' it is applied to. In terms of a higher-order theory of consciousness, this means that the 'semantic' contribution of reflexion[56] to the resulting higher-order thought is always the same, no matter what the object of reflexion is. The semantic contribution consists basically in attributing a thought to oneself. To put it another way, consider a reflexion-expressing sentence "I think that S." In the sense of reflexion, it is a *de dicto* statement: I attribute to myself an object, the thought the content of which is expressed by S. If the semantic contribution of the "I think" is to be identical for all acts of reflexion, it suggests that compositionality applies also to the level of representation of thoughts.[57]

At last, let me bring all the threads together. If we hold that compositionality is a principle that applies not only to linguistic expressions but may, *mutatis mutandis*, apply also to the level of representation of thoughts (the conceptual, as described in 3.1), then we can account for productivity of thinking that is not articulated in any specific language. For otherwise we would either have to claim that all thinking and reasoning (unconscious notwithstanding) is essentially linguistic,[58] or we would have to admit that complex thinking just happens. Assuming the higher-order theory of consciousness, we can thus allow for the intuition that I can be conscious of something without being conscious of its articulation in any language I know. The intuition may be wrong, as I am sometimes inclined to think, but people would still like to claim that they, by introspection, know that some of

[56]Obviously, the contribution may be called 'semantic' only in the derivative sense of semantic contribution of the subsentential part expressing reflexion ("I think") to the meaning of a sentence expressing the higher-order thought itself ("I think S."). Admittedly, such a clarification betrays a question-begging commitment to the claim that reflexion is ultimately a linguistic operation, for it is yet to be demonstrated that language enables reflexion. However, this step could be vindicated if 1) its purpose is taken to be only facilitation of the intuition of what the formality of reflexion means; and 2) if the *functional* contribution of reflexion to our cognition is the same as the linguistic operation of adding the "I think" clause. I believe the latter is true, but it would be lengthy diggression to argue for it here.

[57]I don't think that speaking of a level of representation of thoughts commits us to anything like mentalese. I think, however, that one has to conceive of thoughts at such a level of description where thoughts *represent*. That is, while we may have good reasons to believe that thoughts are actually realized by parallel distributed processes in brain, where it is hard, if not impossible, to imagine what an analogue of compositionality would be, we will still identify thoughts by what they are *about*. Thus at that level of description, we are entitled to treat thoughts as representations.

[58]This is not to suggest that this option is in any way deficient. It may eventually appear to be right, though I reckon that would be for reasons unrelated to compositionality. For the time being, I only want to keep the options open.

their thoughts are explicitly articulated (like inner exclamation "What the hell does that mean?") and others are not.[59] Furthermore, the recursive character of reflexion[60] invites the idea that it obeys the principle of compositionality. While it may be difficult to make sense of compositionality as working for anything else than linguistic expressions, we have to posit it for the conceptual, I think, if we take the argument from productivity and systematicity of *thinking* seriously. I am convinced that the alternative of conflating the conceptual with the linguistic leads to even less desirable difficulties.

4 Meaning and Intentionality

4.1 From consciousness to meaning...

I have argued that the best explanation of the iterative and formal character of reflexion is that it is governed by the same principles as forming sentences of the pattern "I think that S." Sometimes the mystery of consciousness is phrased, in order to make it clearer and tangible, as the problem how one becomes conscious of yet unconscious content. For example, how could computer become conscious (say, of its own states)? Such phrasing may be, however, tendentious, since its presuppositions are by no means uncontroversial. First, is consciousness necessarily a relation of a self to some content – like special index that the content acquires when some internal process is performed? Second, is content conceivable without its apprehension, or in other words, is it meaningful to speak of unconscious content? The phrasing with its presuppositions invites specific theory of mind and meaning, namely representational theory of mind (RTM) coupled with language of thought hypothesis (LOT), the main champion of which is Jerry Fodor. Assuming RTM with LOT, consciousness could be understood as follows (submitted to the constraints of consciousness that I have sketched so far): a creature is conscious of content C, expressible by sentence S, if the creature has currently a token in LOT the content of which is expressed by the sentence "I think that

[59] Whether to reject this intuition, or save it and explicate its sense, depends on what we intend to mean by 'articulation', which again depends on our conception of the link between the linguistic and the conceptual. If articulation is a process in our language-production module, i.e. the process psycholinguistics is interested with, then the intuition should be saved. If, on the other hand, articulation means just conceptualization then the intuition should be rejected. A way to distinguish between these two senses could be auxiliary judgement whether articulation is held to be language-specific, i.e. whether it is identifiable in what language the content is articulated.

[60] For clarification, see appendix A, p. 67.

S."[61] That is, however, more likely an account of the empirical consciousness, which is to be distinguished from the transcendental consciousness pursued here.

The preceding example of an account of consciousness, by means of illustrative RTM that has had a long tradition in the philosophy of mind, was to point at mutual dependence between theory of meaning and theory of mind (and consequently of consciousness). Semantics constrains theory of mind, and vice versa. That is no coincidence, since consciousness is intuitively an effect of conceptualization: whatever we are conscious of, it must be conceptually structured. As for the other direction, any account of meaning more or less directly implies what kind of things the things that *understand* or *act upon* meanings (i.e. minds) must be.

Usually, one does not consider her semantics and theory of mind simultaneously, but selects the starting point at one and then see the consequences to the other. The preference of choice may be based on conviction about the relative conspicousness of what the mind/meaning is, or about primacy in the explanatory or causal order etc. The concept of transcendental self-consciousness is particularly non-commital about theory of meaning, especially if interpreted along the original, highly abstract and formal, Kantian lines. What seems to be clear, however, is that Kant's explication of the transcendental self-consciousness, as the "I think" that must possibly accompany any representation, is not an allusion to the possibility of merely syntactic or formal extension. Taking the explication quite literally as I have suggested drives us to the conclusion that consciousness has perhaps "something to do" with the meaning of the "I think". Perhaps, by exploring what meaning might be, we could get better grip on what transcendental self-consciousness amounts to.

4.2 ... via intentionality...

For most purposes of this work it is not necessary to delve into specifically semantic technicalities; so for convenience and clarity I will make do with discussing options about what is "to have (semantic) content"[62] which is

[61] More elaborate version could also deal with contents that may be expressible by subsentential expressions, such as noun phrases (and these are, allegedly, expressing content, since the semantics of LOT is representational and atomistic), if forms of consciousness other than "I think that *S*" are allowed - e.g. "I see [NP]." Similarly, propositional attitudes other than thinking may be stipulated to be sufficient for consciousness ("I wish that *S*."). However, this can be achieved on pain of consciousness coming in different flavours the common denominator of which remains unexplained.

[62] [Haugeland(1990), p. 384].

reasonably close to having intentionality – according to J. Haugeland, whose article offers a useful overview and classification presented below. As far as I understand it, the concept of 'semantic content', that Haugeland uses to express what intentionality amounts to, can be identified with meaning. So we will eventually come to know what it is for something to have meaning – and that will be just enough for further discussions of consciousness; precise models explaining why something means A rather than B are beyond our needs as long as we have a general idea how the issue can be resolved.

4.2.1 Some general remarks on intentionality

i. Some things can be said to have meaning only derivatively (for instance, thoughts confer meaning to sentences that express them, thus sentences have derivative intentionality and thoughts original one), which presupposes there be something exhibiting original intentionality, which is the actual matter of discussion.

(ii.) Intentionality as a relation between the content bearer (e.g. a thought) and what it is about (e.g. a state of affairs) is peculiar in that the latter may not exist/occur. Yet intentionality also ought to have causal powers in order to make sense of our folk-psychological explanations that we did *this* for *that* reason.

iii. All three accounts of intentionality to be presented are designed to be compatible with materialism, which specifically precludes resorting to stipulation of intentionality as an intrinsic property (of things considered contentful).

(iv.) Consequently, since content (meaning) of something cannot be determined only by its own properties, it must be determined by its relation to other things, by "some larger pattern into which it fits."[63]

Points (ii.) and (iv.) are put in brackets because the sense Haugeland attaches to them seems to be biased in favour of his own, pragmatic account of intentionality. The non-existence of the second relatum mentioned in (ii.) is interpreted as necessarily a normative feature of intentionality (e.g. a belief can be right or wrong). The dependence on some larger pattern in (iv.) is intepreted as holism. I don't think Fodor, for example, would concede to these interpretations, but since these are important points that would come into consideration anyway, I keep them in the list.

[63] [Haugeland(1990), p. 386].

4.2.2 Fodor and neo-cartesianism

Neo-cartesianism is a label for the doctrine that only mental states are in the proper sense intentional. Mental states are some cognitive system's internal tokens that can be specified both intentionally (as symbols) and formally (as, for example, physical configurations). Thus some clearly delineated physical thing works as a symbol with equally determinate meaning (so-called token-token identity theory). Meaning of a symbol depends entirely on its formal properties, or as Haugeland puts it more precisely: "the semantics of the symbols is a systematic function of their formal character as tokens."[64]

In line with the general point (iv.) from the previous paragraph, Haugeland interprets neo-cartesianism as endorsing that "separate mental states have their contents in virtue of their systematic relations to *one another*."[65] Though it is certainly a reasonable position, Fodor does not seem to take it. In his recent monography on language of thought, he puts forward purely referential, atomistic semantics that can supposedly do justice to intentionality. Specifically, one of the tenets of his theory reads:

> The metaphysics of concept possession is atomistic. In principle, one might have any one concept without having any of the others (except that having a complex concept requires having its constituent concepts).[66]

This can be so because Fodor builds up his theory of mind as representational and semantics as referential. Neither reference nor representation is a normative relation; for a symbol to refer to something is to be *locked to* it: "If M is a mental representation locked to property P, then 'tokens of P cause tokens of M' is counterfactual supporting."[67] So stated, reference/locking is a nomological relation between a symbol and a property, but its naturalization is straightforward: just find the actual causal mechanism that reliably correlates internal states with properties.

Fodor tries to evade the charge of the necessarily normative character of meaning/intentionality. He seems to concede that symbols can be used correctly or incorrectly in speech acts (so he accepts the normative character of symbols-in-use) and also that this normativity cannot be reduced to causation (which is, however, sufficient for an account of reference), but he replies:

---

[64] [Haugeland(1990), p. 389].
[65] [Haugeland(1990), p. 388].
[66] [Fodor(2008), p. 141].
[67] Ibid.

> [W]e're committed to LOT; and LOT, though it is a system of representations, isn't a system of representation that anybody *uses*, correctly or otherwise. One doesn't use thoughts, one just has them.[68] [Fodor(2008), p. 203]

Coming back to Haugeland, he insists that "specifying [tokens'] symbolic contents can be regarded as interpreting them."[69] The interpretation in neocartesianism is then constrained by the demand for 'truth' (correspondence of the pattern of tokens to some state of affairs) and 'intrasystematic rationality'. But even if interpretation of symbols has to take into account the holistic pattern of tokens and corresponding referents in order for the contents to be both systematic and sensible,[70] it does not provide the criteria for distinguishing original and derivative intentionality. A piece of text may exhibit the same interpretation-relevant pattern as a set of mental states of a thinking thing. Thus what makes the latter the domain of original intentionality are causal relations between mental states *qua* inner symbols. Thanks to the token-token identity, the very same thing that is the basic unit of semantic evaluation (that means: is a content bearer) has causal powers due to its physical properties. It is because the appropriate tokens (e.g. such that their contents are A and $A \rightarrow B$) actually jointly *cause* tokening of the right symbol (B) that the system of so related tokens exhibits original intentionality. This is an expression of computational theory of mind (CTM) which is so intimately related to LOT that either can be regarded as the motivation for the other. Fodor, at least, builds his LOT semantics with the goal of getting CTM work.

Ultimately, symbol's content is determined by its formal (syntactical, to be more specific) properties because these alone determine its causal powers; and the content of a symbol depends entirely on causal relation to the world and/or other symbols.[71] Fodor realizes that his emphasis on reference as the sole and basic semantic relation leads to difficulties in accounting for many of the subtle differences in meaning which we recognize in extensionally

[68] The commitment to LOT stems from the indispensibility of folk-psychology in our everyday life. Fodor holds that the ontological commitment of belief or thought ascriptions had better be taken seriously; not as mere *façons de parler*.

[69] [Haugeland(1990), p. 389].

[70] Systematicity implies that any type of formal tokens corresponds to a type of content, e.g. that any token of the string-type "c-a-t" has token of CAT as its content. That the interpretation be sensible means that the to-be-interpreted pattern of tokens should correspond to the pattern of things the tokens are about, e.g. refer to. The details are at [Haugeland(1990), p. 391].

[71] "You connect the causal properties of a symbol with its semantic properties *via its syntax*. The syntax of a symbol is one of its higher-order physical properties." [Fodor(1987), p. 18].

equivalent expressions. Unsurprisingly then, large part of his LOT 2 book is devoted to the attempt to deal with Frege's problem that gave rise to the familiar distinction between sense and reference. It is enough to say that LOT purportedly meets Frege's problem by an appeal to CTM from which it follows that there could be more types of representations for the same referent.[72] The direction of looking at meaning Fodor maintains is more important: see how mental particulars bear content and how thinking can be a causal process; the rest should follow.

Why does Fodor try so stubbornly to avoid semantic holism and sticks to atomism? Because he is convinced that "compositionality is at the heart of the productivity and systematicity of thoughts."[73] Most of the semantic theories discussed in his LOT 2 are criticised on compositionality's behalf. Any semantics that fails to be compositional should be rejected outright. The architecture of his theory of meaning follows from the demand for compositionality:

> Referentialism must be right about the content of intentional states because compositionality demands it; atomism must be right about the individuation of concepts because compositionality demands it; and thought must have constituent structures because compositionality demands that too. ... Most of what we know about concepts follows from the compositionality of thoughts.[74]

Apparently, compositionality according to Fodor is a fact, to refer to the distinction made in 3.2.1.

4.2.3 Dennett and neo-behaviourism

As the reference to behaviourism suggests, the major point of divide from neo-cartesianism is that beliefs, thoughts, reasons and other intentional entities are not held to be mental particulars. The 'neo-' distinction is allegedly earned by somewhat more realistic attitude to the intentional ascription compared to the old-style behaviourism. The pattern relevant for content de-

[72] It means, roughly, that mental state types that both refer to Cicero may have different physical properties so that one can be interpreted as the concept CICERO while the other as TULLY. As Fodor summarizes, "From CTM's perspective, the existence of Frege's problems shows *at most* that reference isn't sufficient for the individuation of concepts; something further is required. But Frege's problem *doesn't* show that the 'something' else is a parameter of content; for example, that it is something like a sense." [Fodor(2008), p. 70].

[73] [Fodor(2008), p. 20].

[74] Ibid.

termination is the pattern "in the interactions between the system and its environment."[75] An interaction is *necessarily* described in intentional terms as a pair of perception and action.[76]

How do mental (or preferably 'inner', or 'cognitive') states come into scene? By the need to ascribe some intermediary states to the system in order to explain the system's behaviour. Thus, a chess machine can be ascribed states of knowing the chess rules and current board configuration, aiming to win, strategy to control the centre *ceteris paribus* etc. Most importantly, however, a particular state may have no other determination than intentional,[77] and its ascription "depends exclusively on the pattern exhibited in perceptions and action."[78]

Intentional ascription is guided by the aim to "maximize a system's overall competence."[79] In other words, the ascription should be such that the manifest behaviour is rational with regard to them, and successful in normal circumstances.[80] The trouble is that in maximizing a system's competence we may ascribe too much of intentionality.[81] Somewhat obscure way out is proposed, namely that one ought to rationalize only the *manifest* behaviour, but the problem remains, I think, if only because what behaviour is manifest is again observer-relative, and thus it seems to come down to a version of Occam's razor for intentional ascription: "Do not multiply intervenient cognitive states beyond necessity."

In an attempt to naturalize the intention-ascription process, Dennett ex-

[75] [Haugeland(1990), p. 395].

[76] The modal qualification is appropriate because to get started with some such pattern at all we need to have pairs like ['sees a lion';'runs away'] rather than ['part of retinal image is caused by a lion positioned north-east';'runs south-west']. Without interpreting the behaviour of a system at this basic level as actions (not events), we could not say whether the interaction pair to be included in the relevant pattern is not, by chance, ['patch of blue sky is represented in the retinal image';'runs south-west']. In consequence, the relevant pattern is observer-relative, though constrained by the demand for consistency and predictive success. That is, after all, what Dennett's intentional stance implies.

[77] That is to say that such cognitive states may not be so clearly delineated at the level of their physical realization. For this reason, Haugeland remarks, neo-behaviourism naturally leans to the connectionist branch of cognitive science, while neo-cartesianism is in accord with the classical AI.

[78] [Haugeland(1990), p. 396].

[79] [Haugeland(1990), p. 398].

[80] Unfortunately, specifying what counts as *normal* circumstances of an action already presupposes knowledge of what the action should achieve, which is part of what is to be ascribed. Perhaps we can replace it vaguely with 'most of the time'. The idea is quite simple in the end: don't say cats want to make ratcatchers lose their job by catching mice; that would render them stupid – because hopelessly unsuccessful.

[81] Haugeland uses an illustrative example of a mousetrap that wants to kill mice, believes it can do so when they touch the bait, decides to snap etc.

plains it as a capacity which evolved under the pressure of the need to predict ever more complex behaviour of co-evolving organisms, including ourselves (that prediction is a favourable skill itself depends on other ecological facts, such that any organism is either a prey or a hunter of some other). Therefore, in Dennett's view, "[t]he only 'original' intentionality anywhere is the mere *as-if* intentionality of the process of natural selection viewed from the intentional stance."[82]

Predictive success is then a normative measure of correctness of intentional ascription. That it is not the only criterion of correctness seems to follow from the possibility of many predictively equal interpretations some of which would be rejected by common sense (see Haugeland's mousetrap example). Dennett addresses this issue when he notes that in interpreting a behaviourally simpler system than us (which is most of the time) our language *cuts too fine*. When *we* conceptualize the intentions to be ascribed, we tend to express them in propositions that imply too fine a distinction in the system's cognition.

> For instance, when a frog's tongue darts out and catches whatever is flying by, the frog may make a mistake - it may ingest a ball bearing thrown by a mischievous child, or a fisherman's lure on a monofilament thread, or some other inedible anomaly. The frog has made a mistake, but *exactly* which mistake(s) has it made? What did the frog "think" it was grabbing? A fly? Airborne food? A moving dark convexity? [Dennett(1996), pp. 50-51]

The tendency to specify the intension[83] of a system's intention can be misleading. "The misguided goal propositional precision", as Dennett calls it, is the greatest danger of the intentional stance. If we are to avoid it, we need to realize the following:

> If the agent under examination doesn't conceive of its circumstances with the aid of a language capable of making certain distinctions, the superb resolving power of our language can't be harnessed directly to the task of *expressing* the particular thoughts, or ways of thinking, or variaties of sensitivity, of that agent. (Indirectly, however, language can be used to *describe* those particularities in whatever detail the theoretical context demands.) [Dennett(1996), p. 55]

[82] [Dennett(2006)]

[83] As Dennett notes, the intentional stance is referentially opaque, i.e. the plausibility of intentional ascription depends on *how* the extension of an intention is picked up.

It is the difference between *expression* and *description* which makes Dennett's approach viable and theoretically profound, at least from the perspective of cognitive science.

Finally, meaning explicitly comes on stage when lingustic interactions can be discerned among others. And they can by virtue of relative idenpendence of other subject-environment interactions.

> There could, for instance, be cognitive states of a distinctive sort that are already formulated verbally within the linguistic faculty - verbal cognitions, we might call them. And dispositions specific to these states might interact among themselves in a way that is relatively divorced from the agent's other dispositions, and more narrowly "logical." [Haugeland(1990), p. 402]

Linguistic competence is still one of many competences a system might exhibit and so meaning of a speech act is determined by the same principle of ascription whose main criterion is success. Thus meaning of an utterance comes from speaker's intention - the change in environment she wants to achieve by this action. Gricean conversational implicatures, for example, exploit this point explicitly. Still, what remains to be seen, from the semantic point of view, is what determines meaning of words and sentences (i.e. linguistic types, not tokens in context as utterances and speech acts).

4.2.4 Brandom and neo-pragmatism

Neo-pragmatism thinks of anything contentful essentially in terms of social practices. Unlike in neo-behaviourism, the pattern relevant for content determination cannot be formed by *any* interaction between a subject and its environment but mainly (or perhaps only) by interaction that can be shaped by, and subjected to, 'censouriousness'. Censorious behaviour is essentially that which promotes conformism - the general tendency to act in accord with already established norms. Thus punishing deviations from norms and rewarding their abidance is censorious in that it reinforces the behavioural pattern which fits in the present normative structure, way of life, or culture, if you will.

Haugeland, who himself sides with neo-pragmatism, tries to delineate conformism so that it is a behavioural trait which only acting creatures can exhibit.[84] Note, however, that subject's environment is in a sense censorious as well, namely in the sense of success evoked in neo-behaviourism. Sickness

[84] "*Conformism* here means not just imitativeness (monkey see, monkey do), but also censoriousness - that is, a positive tendency to see that one's neighbors do likewise, and to suppress variation." [Haugeland(1990), p. 404].

caused by eating wildly coloured mushrooms is a punishment that can be said to be promoting conformism to the norm "Don't eat wildly coloured mushrooms."[85] It does not help appealing to the intuition that censoriousness is surely to be understood as intentional (and the Nature does not clearly harbour any such explicit intentions). For it not only begs the question of what constitutes intentionality, but it also contradicts some obvious cases of censorious behaviour that is not manifestly intentional. For example, our unconsciously reserved behaviour towards frowning people (and warm one towards smiling people) is perhaps a great enough punishment (reward) to promote the social norm of smiling *ceteris paribus*.[86] Thus the crucial pragmatist restriction that the normative can only emerge from *social* behaviour presupposes a theory of action such that 1) acts towards others may be performed on the basis of only implicitly 'recognized' means and ends; and 2) acts towards others ought to be distinguished from actions aimed to bring about a change in environment.

At the bottom level, conformism can be a hard-wired trait of some organisms (or systems generally) - such that the species is in effect called social. A closer look at normativity of linguistic practices may reveal a feature both distinguishing neo-pragmatism from neo-behaviourism and vindicating the emphasis on social interaction. As R. Millikan observes, linguistic conventions sustain because conforming to them helps to solve a coordination problem.[87] Thus far it is just a variation on the neo-behavourist success story. However, linguistic convention, unlike practical conventions like eating with knife and fork or chopsticks, are established by there occuring a precedens which is at that time contingent in respect to the function it serves and which is followed from then on thanks to conformism. To illustrate, at the time somebody first called a dog "chien" (*and* in such circumstances that others' conformist nature made them follow) there was nothing constraining that first call to be "chien" rather than "pes" or "london" (unless these were

[85] A similar point is made by Dennett in his reaction to Brandom's *Making it Explicit*. Dennett disagrees with Brandom on the matter whether norms pertain essentially to communities only. See [Dennett(2006)].

[86] Or is that a norm only in a metaphorical sense, not the norm proper? Taking such a stance would, from the pragmatist point of view, betray unwarranted bias to some prerequisite of norm-institution (e.g. language or consciousness). If I understand it correctly, pragmatism hopes to explain how any such putative prerequisite could arise as a cultural phenomenon by virtue of norms, the presence of which does not necessarily depend on the phenomenon itself.

[87] "A coordination problem arises when people have a purpose in common which must be achieved by joint action, where the contribution that each must make will vary depending on what each of the others contributes, and where there is more than one acceptable way of combining contributions to produce a successful outcome." [Millikan(2005), p. 55].

already in use, of course). Importantly, others will comply with the precedent use only if it is also advantageous to them as hearers. If "chien" were being used inconsistently in presence of dogs, cats, birds etc., it would not be taken upon (unless one could distinguish a different consistent usage in it, e.g. that of "vertebrae"). Similarly, if promises were broken more often than not and with no appropriate sanction, the institution of promise would lose its binding force and therefore cease to exist.

> A corollary is that the functions of public-language forms are not on the same level as either speaker purposes or hearer purposes taken alone. The conventional functions of language forms are not, for example, merely standard speaker purposes. Conventional language forms are selected for performing services satisfactory at once to both partners in communication. Their functions must balance speaker with hearer interests. Because the conventional function of a linguistc form will remain stable only if it continues to serve the interest of both speakers and hearers often enough, I call it a 'stabilizing function'. Linguistic 'meaning' in the sense of stabilizing function is on an entirely different level from, for example, average speaker meaning. [Millikan(2005), p. 58]

The last quoted sentence provides us with means for distinguishing between the general, public meaning of a linguistic form and meaning in the sense of speaker's intention; a distinction that is presumably difficult to articulate in neo-behaviourism, for there the latter is the sole sense of 'meaning'.

The concept of norm is obviously central to pragmatism. What exactly are they then?

> When behavioral dispositions aggregate under the force of conformism, it isn't herds that coalesce, but *norms*. ... Like herds, norms are a kind of "emergent" entity, with an identity and life of their own, over and above that of their constituents. [Haugeland(1990), p. 405]

For an outsider, such a claim may seem to be a great ontological commitment; for pragmatists, however, an inevitable one.

As I have already hinted above, if neo-pragmatism is to provide a causal story of the origin of intentionality, norm-abiding must not require explicit rule-following. This point should be granted unproblematically if the biological or evolutionary account of conformism, outlined above, is accepted. In an important footnote, Haugeland clarifies the ontological commitment of neo-pragmatism:

> There is a natural extension of the notion of following an explicit rule to that of a computer following an explicit program; and it might be held that no material system could in fact exhibit the required behavioral complexity and versatility unless it explicitly followed a complex rule (program) "internally." But this is a strong and separate claim that calls for an independent argument. The question is presumably in some sense empirical. [Neo-pragmatism], like [neo-behaviourism], need not be concerned with how the dispositions are "implemented"; hence, it is also entirely compatible with "distributed representation" models of how it all works. [Haugeland(1990), footnote 24, p. 422]

Intrinsic to norm-abiding is the right recognition of circumstances in which the norm applies. I may know that the younger should be, *ceteris paribus*, introduced to the older first and may wish to comply, but I still may violate the norm by failing to recognize who is younger. Loosely speaking, part of the content of a norm are circumstances of appropriate application, which is just to say that "norms have a kind of 'if-then' character, connecting sorts of circumstance to sorts of behavior."[88] Consequently, norms are interrelated by their conditions of right application.

Finally, what is the neo-pragmatist account of language? Unsurprisingly, pragmatism advises to consider linguistic forms as tools. As hammers can be used in a wrong way (holding it upside-down) or for wrong purposes (killing), so similarly words or sentences may be used incorrectly or abused (for lying, for instance). The reason why linguistic expressions are specific tools among others is that their primary purpose is "to affect the current normative circumstance, including the status of both utterer and audience."[89] For example, making a move in chess affects the normative circumstance in the sense that now it is the other player's turn, configuration of the board is different etc. Similarly, asking someone a favour passes the initiative to the trustee and binds her to either comply and earn one's gratitude or reject on pain of severing the personal relation.[90] Original intentionality then belongs to linguistic forms as conceived in their public, norm-constituting use.

R. Brandom, the apostle of inferentialism, puts forward a more complicated view of the linguistic intentionality. He points out that where carte-

[88] [Haugeland(1990), p. 408].

[89] Ibid., p. 411.

[90] As the example shows, some moves in language games affect or depend on extralinguistic circumstances; unlike in case of chess that is in this sense self-contained. As M. Lance likes to put it: language is more like baseball than chess in that it requires a pitch, a baseball, etc. (not to mention obtaining physical laws). For more detail see [Haugeland(1990), p. 412].

sianism and traditional semantics relegate explanations on *representation*, pragmatism will naturally turn to *expression* (in the sense of process by which "inner becomes outer", as in the simple case of a gesture expressing a feeling). Brandom's preferred word is explicitation, that is "making explicit what is implicit." While he claims that explicitation is conceptualization of some subject matter, he immediately adds that the implicit may not be independent of its ways of explicitation:

> [S]pecification of what is implicit may depend on the possibility of making it explicit. And the explicit may not be specifiable apart from consideration of what is made explicit. On such a view, what is expressed must be understood in terms of the possibility of expression. Such a *relational* expressivism will understand linguistic performances and the intentional states they express each as essential elements in a whole that is intelligible only in terms of their relation.[91] [Brandom(2001), p. 9]

This sounds very Kantian indeed. Relational expressivism might remind us of the famous dictum that "thoughts without content are empty, intuitions without concepts are blind." However, it seems to be a hindrance in the quest for explaining intentionality in naturalist framework, for the intention-expression interdependence forms an unbreakable circle.[92]

Inferentialism understands meaning of a linguistic expression as consisting of rules of inference.

> Grasping the *concept* that is applied in such a making explicit is mastering its *inferential* use: knowing (in the practical sense of being able to distinguish, a kind of knowing *how*) what else one would be committing oneself to by applying the concept, what would entitle one to do so, and what would preclude such entitlement. [Brandom(2001), p. 11]

Importantly, the rules of inference can themselves be made explicit by specific linguistic means. That, according to inferentialism, is the role of logic.

[91] The first occurence of 'specification' in the quotation is to be read very loosely, I think, otherwise the claim would be trivial, since every specification (in the strict sense of articulating the differentiae of the species belonging to a genus) is explicit. The question is what the loose sense is supposed to be. Determination, perhaps?

[92] No wonder that to break the circle with a one-way oriented explanation, Brandom resorts to pragmatist's ultimate appeal: an evolutionary story about the origin of normative social practice from merely differentially responsive creatures' acts. Sellars offers one such story in his *Empiricism and the Philosophy of Mind* (recounted here on p. 53).

Saying "If something is a man, then it is a mammal." is making explicit the inferential rule governing the correct use of "man"; it says what we are doing when saying that someone is a man. Interestingly enough, Brandom describes such using of logic as promoting "semantic self-consciousness". Whether this concept could shed any light on transcendental self-consciousness pursued here remains to be explored.

4.3 ... and back again

Now that we have considered various accounts of intentionality, we can turn our focus back to consciousness.

4.3.1 Fodor and computation

Let's start with Fodor. His LOT hypothesis coupled with CTM might seem attractive for anyone trying to provide a naturalist account of mind or consciousness. For it gives a clear idea how the linguistic capacity can be *constructed*. If Fodor is right, then computers can in principle speak (with all the intricacies traditionally associated with full-fledged language). For those who believe that consciousness has something to do with language, it is a good reason to hope that if Fodor's account is correct, it may also contribute to the explanation of consciousness. Though I am going to argue that Fodor and neo-cartesianism do not offer the means to explain transcendental self-consciousness, I do grant that his project can be easily empirically evaluated - were we presented with a machine supposedly able to talk, we know how to verify that.[93]

Fodor's metaphysics of mind is quite simple, so we can go through all alternative theories of consciousness it may allow for. Minds consists essentially of mental states which are relations to symbols - mental particulars. Let me illustrate it by Fodor's account of propositional attitudes: to believe that P is simply to have P in one's "belief box", which, metaphysically speaking, means bearing the relation of believing to a proposition.[94] Mental processes are causal sequences of mental tokens - particular symbols. Fodor believes that his account has the virtue of explaining, among other things, the possibility of practical syllogisms. If I have P in my "wanting box" and P *iff* Q in my "belief box", then I act to bring it about that Q. A computation yielding Q into "to-do box" could only be systematically performed if there

[93] I will develop the motive of Turing test further below.
[94] Cf. [Fodor(2008), p. 69] or [Fodor(1987), p. 17].

is the same intentional object present in my wanting box and belief box (i.e. P).[95]

Could consciousness amount to having some content in "consciousness box"? Or, to express it in less controversial terms, could consciousness be a relation to mental tokens, like a special propositional attitude? Yes, but only in a sense which will not explain what consciousness is, how it arises, what it requires etc. Consciousness, in the *objective* sense, is always *about* some content (hence objective); thus far, saying that consciousness is a relation to mental tokens is a trivial statement. But that expresses only what we have identified, thanks to Kant, as empirical consciousness. For Fodor's account to have any relevance for transcendental consciousness, the actual computation that puts something into consciousness box would have to be explored. For the rules *behind* this consciousness-creating computation are what specifies the condition of possibility of (empirical) consciousness, and thus they would be the closest Fodorian analogue to transcendental consciousness.

Such an interpretation of transcendental consciousness, however, has gone quite far from the more literal one I have envisaged. So let me consider whether the syntactical operation of adding "I think" to P could tell us anything about the consciousness-creating computation. Meaning of the "I think" of transcendental consciousness is specific in that it does not make sense to *doubt* whether I am actually thinking that P or not. Wittgenstein draws on the same point when he argues that "I know" in "I know I am in pain." is an idle wheel. However, it does make sense to doubt whether someone else thinks that P or whether I am six feet tall. Therefore the "I" and "think" of transcendental consciousness must have quite different sense than in their ordinary use. But Fodor's atomistic account with well-behaving compositionality precludes such an option. Certainly, one could think of a special symbol type whose role would be just this "I think" that can accompany any mental state; but that would not explain anything. Perhaps, the idea that consciousness could occur in virtue of a symbol added to a proposition seems preposterous right from the beginning, and I do not imply that Fodor might have ever considered it as an explanation. I only wanted to expose how limited is Fodor's theory in means to tackle consciousness. In the end, he probably regards the problem of consciousness insolvable:

> Tom Nagel once wrote that consciousness is what makes the philosophy of mind so hard. That's almost right. In fact, it's intentionality that makes the philosophy of mind so hard; consciousness is what makes it impossible. [Fodor(2008), p. 22]

[95] See [Fodor(2008), p. 15].

Even though I have just argued to the conclusion that the computational theory of mind built up on LOT fails to provide insights to transcendental consciousness, it ought not to be inferred that artificial systems like computers cannot in principle be conscious. I believe that consciousness is an empirical matter and consequently that whether something is conscious or not can be empirically assessed. The assessment does not require investigation of internal processes of the thing, after all we deem other people conscious without checking whether they really have brains.[96] Hence, the idea behind the Turing test can be extrapolated from thinking to consciousness.[97]

Just to make things clear, the claim that consciousness is an empirical matter does not imply, I believe, that *transcendental* consciousness is a physical feature nor that it is a behavioural aspect. As I have argued at page 23, transcendental consciousness specifies certain 'logical' characteristics of our understanding that must hold in order for the empirical consciousness to be possible. We may recognize something as conscious even though we do not know what makes it conscious. What's more: we could construct a conscious thing without being aware of what constitutes the *élan cognitive*.

4.3.2 Dennett, Sellars and their stories

What can we learn about consciousness from neo-behaviourism in general and Dennett in particular? The emphasis on interactions between the subject and its environment can supply an account of the selective pressures under which consciousness has evolved. Understanding the purpose of consciousness, however, may shed only little light on what constitutes consciousness.

Apart from the evolutionary perspective, the subject-environment interaction may have a more direct consequence for consciousness, as Dennett expounds in his *Consciousness Explained*. There he develops his *multiple drafts model*, the core of which is that conscious contents are indeterminate

[96] However, one could object that looking in other people's heads is unnecessary because their having consciousness is a *status quo*, a null hypothesis of folk psychology, so to speak. Therefore, we may be able to empirically falsify the null-hypothesis, but never to verify it, which is as good as real scientific approach. Yet this 'agnostic solipsism' may allow that some 'systems' could be assigned consciousness by default *if they share with us* what is thought of as indicative of consciousness, e.g. the form of life, linguistic capacity etc. I am convinced that whatever the indicators are, they manifest behaviourally.

[97] Turing designed his test as a way of deciding whether machines can *think*, not whether they are conscious. But if consciousness manifest overtly and if its presence is not independent of its manifestation, then we can imagine how the Turing test could answer even the query about consciousness. Antecedent of this conditional ultimately turns to the famous question whether zombies are metaphysically possible (or, to list the reasonable alternatives, inconceivable *sensu stricto*, or materially impossible). For the Turing test idea, see [Turing(1950)].

until the subject is required to make a specific response to environment. What we are conscious of is not settled until it is expressed. "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)."[98] In the end, Dennett attributes special role to linguistic expression among the effects that determine conscious contents:

> 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(1996), pp. 205-206]

Again, there is no reality of consciousness besides its effects that can manifest overtly - which is after all what we would expect from neo-behaviourism. There is a flow of information[99] supplied by various discriminators in the brain. The flow is continuous and undergoes constant revision (hence *multiple drafts* model).

Dennett realizes he ought to explain the introspective finding that we do have conscious thoughts which nevertheless elicit no clear manifestation. He can offer two replies. First, mental contents (i.e. information, in his understanding) are conscious to certain degree - there is no non-arbitrary way to decide what is to count as conscious because consciousness is conceived of in terms of disposition to (behavioural) expression, or "the competition for the control of behaviour" in the quotation above. Second, he invokes a story about internalization of the process of self-probing through language. Originally, one used language to ask and respond to others (for information, favours etc.). Eventually, one asked and responded herself to her own benefit which reinforced the whole process so that it became a habit of inner speech that we now take to be the stage of consciousness.[100]

[98] [Dennett(1991), p. 132].

[99] Dennett is particularly careful to avoid any claim that would support the interpretation along LOT lines. Thus speaking of information on his part does not presuppose any medium of representation; the content of information is specified functionally.

[100] Cf. [Dennett(1991), p. 195], [Dennett(1996), pp. 195-201].

Another contribution of the neo-behaviourist paradigm is the possible significance of the intentional stance to consciousness. If being conscious of something amounts to ascribing a thought to oneself then we could surmise consciousness originates in adopting the intentional stance to oneself. Intuitively, however, conscious thoughts differ significantly from thoughts ascribed to other subjects. The difference consists not only in the intrinsic information that conscious thoughts are *mine*, not someone elses, but more importantly in that the subject does not have to observe her behaviour in order to realize what she is conscious of. In other words, the self-ascription of thoughts is non-inferential. Yet if we buy the premise that the intentional stance is a predecessor of consciousness (at least in the order of explanation), we need to understand how the inferential process could become cognitively direct.

Sellars in his (1956) offers an explanation by telling a story similar but more complex than the Dennett's story outlined above. We are to imagine a community whose members use language to ask favours, express their own needs and even to make one's intention-ascriptions explicit (e.g. by saying "The man over there is hungry."). It so happens that someone (called Jones in the story), in observing the intelligent behaviour of his fellows, "develops a *theory* according to which overt utterances are but the culmination of a process which begins with certain inner episodes."[101] Thus folk-psychology is invented in which thoughts, beliefs, desires etc. are theoretical entities hypothesised to be inner episodes modelled on language (hence the inner speech metaphor). So far the story has not brought anything new to the intentional stance paradigm, except for emphasizing the *theoretical* status of intentional vocabulary, which *ipso facto* implies it could have originated only through language.[102] So, what is needed in order for the folk-psychological vocabulary, that initially had only theoretical status, to gain a reporting role? In other words, how does Jones acquire the capacity to report (non-inferentially) on his occurent thinkings? If the habit of ascription of the theoretical entities gets established in the community, then one can learn to report on her beliefs by the trial and error method, provided others will correct her guesses if the behavioural evidence is to the contrary.[103] So it is the others who posess the original authority of what I believe. This

[101] [Sellars(1956), §56].

[102] Sellars insists that "the intentionality of *thoughts* can be traced to application of semantical categories to overt verbal performances." [Sellars(1956), §50] This can be considered as a reason for classifying this story under the neo-pragmatist label to which Sellars naturally fits, according to the historical interpretation. While labels don't matter, whether Sellars story is related to the intentional stance does.

[103] Cf. [Sellars(1956), §59].

counterintuitive result can be mitigated by the following reasoning: it is not until one learns the basics of self-ascription of thoughts that one can indulge in the complex stream of thoughts (inner speech) on which she is granted the full authority. Again, much depends on whether we take the fable to express a historical story, a stage in children's mind's development or elucidation of logical relations between mind and language.

4.3.3 Brandom and normative scorekeeping

When reading Brandom, one gets the impression that it cannot be stressed too much how normativity is important to meaning and intentionality. As Sellars pointed out, one way to understand what normativity of meaning amounts to is to consider what sentences *about meaning* actually intend to express. Saying that "dog" means dog, the four-legged best friend of man, is not to state a *fact* about meaning but to indicate what the *proper* use of the word is.[104] The underlying idea is more subtle than the favourite Wittgensteinian line that "the meaning of a word is its use in the language."[105] For the fundamental, basic thing endowed with meaning is a *judgement* - an act of *endorsing* a claim. Brandom employs the distinction between endorsing and entertaining (a thought, representation etc.), where the latter is supposed to be an outcome of "subtracting" commitments from the corresponding judgement.[106] He notes he thereby follows the work of Kant, for whom the judgement was the basic act of understanding and a "minimal unit of responsibility", and Frege, for whom taking a sentence to be true is a prerequisite for the derivation of meaning via context principle.[107] So, to mimic the Wittgensteinian line, meaning of a word is its *correct* use in the language.[108]

[104]The example with dog may be, despite its simplicity, obscuring since one could read off the normative force as being tightly related to reference: do not use "dog" in a way that it could not refer to dogs if your judgement is to be true. Perhaps meaning of words like "fun", "virtue" or "dull" would serve as better examples in this respect.

[105][Wittgenstein(1953), §43].

[106]Perhaps a clearer insight is provided by the following interpretation of Frege by Brandom: "For him, merely entertaining a proposition is just endorsing various conditionals in which it appears as antecedent or consequent-and thereby exploring the circumstances under which it would be true, and the consequences that would ensue were it true."[Brandom(2006), section 2].

[107]Cf. [Brandom(2006), section 5].

[108]Obviously, words can be *abused* (and yet contribute to a meaningful expression) in respect to their correct meaning, but in that case, the meaning is likely to be parasitic on the correct use - unlike in case of a *mistake* when simply a wrong word has been chosen. Naturally, as language evolves, former cases of abuse can be incorporated as a standard at next stage.

Brandom explicates concepts, in line with the Kantian paradigm, as "rules that express what is a *reason* for what."[109] Thus concepts determine what one is committed to when endorsing a judgement articulated by them. Furthermore, while some concepts exhibit a recognitional, non-inferential aspect (e.g. concepts pertaining more or less directly to sensation), this can be so only thanks to other concepts being inferentially related to them, which seems to be a consequence of denouncing the Myth of the given. Now, being aware of something is bringing it under a concept, thus making a judgement.[110] And undertaking the responsibility for the judgement presupposes one keeps track of what one has committed herself to. In order to take responsibility, make *binding* claims, I have to relate the binding force to the transcendental subject where the commitments are united. Thus the unity of apperception is qualified as *transcendental* just for this reason that making a judgement (apperceiving) presupposes a subject that undertakes the commitments. In Brandom's words, "The transcendental unity of apperception is 'transcendental' because the sorting of endorsements into co-responsibility classes is a basic condition of the normative significance of commitments."[111]

The requirement for normative scorekeeping on the subject's part suggests that one ought to be able to realize what commitment a particular judgement entails; or, in Brandom's parlance, make the commitments explicit. This is enabled by the semantic discourse about meanings.

> [B]y using *logical* vocabulary, I can make explicit the implicit inferential commitments that articulate the content of the concepts I apply in making ordinary explicit claims. Here the orginal inferential-propositional model of awareness (in the sense of sapience) is applied at a higher-level. In the first application, we get an account of *consciousness* - for example, *that* Leo is a lion. In the second application we get an account of a kind of semantic *self*-consciousness. For in this way we begin to *say* what we are *doing* in *saying* that Leo is a lion.[Brandom(2001), p. 20]

A new thread unfolds here: consciousness as awareness of the *meaning* of a judgement. Intuitively, the idea makes little sense as we in our conscious thinking are not aware, by introspection, of being preoccupied with commitments that our thoughts entail. So what sense can we make of the idea?

[109] [Brandom(2006), section 6].
[110] Cf. [Brandom(2001), p. 16].
[111] [Brandom(2006), section 5].

We could consider the following. By introspection, we may concede that the paradigmatic conscious act is the stream of thoughts commonly referred to as inner dialogue. The dialogic form consists in drawing inferences, arguing against some inferential moves to the consequence of changing a premise, withholding the conclusion, etc. All the moves in the dialogue can be regarded as motivated either by semantic considerations ("A actually means that I cannot hold both B and C") or by introducing a premise taken for granted ("I am convinced that A, so what follows is..."). Semantic considerations help us to decide what judgement we should actually endorse by disclosing the related commitments, some of which may be worth avoiding. Since committing oneself is a socially significant act, it is beneficial for the subject to make an informed choice. That is, before speaking one's mind, one should better mind what she's going to say.

Introspective analyses may perhaps point to other conscious experiences which do not invole any explicit semantic considerations and could be ascribed *implicit* considerations only with a great stretch of philosophical imagination. This work, however, is not intended to do justice to all *species* of conscious experience but to consider features of meaning that may turn out relevant for transcendental self-consciousness. Now that the features have been collected, classified and described, we may proceed to the final stand.

5 Conditions of possibility of speaking

Having discussed some features of language related to cognition and consciousness in general, what can we surmise specifically about the relation between language and transcendental consciousness, which has been identified as the core of the mind-body problem?

First, let's discuss the purported unity of consciousness. Consider the summary by A. Brook:

> For Kant, consciousness being unified is a central feature of the mind, our kind of mind at any rate. In fact, being a single integrated group of experiences (roughly, one person's experiences) requires two kinds of unity.
>
> 1. The experiences must have a single common subject.
> 2. The consciousness that this subject has of represented objects and/or representations must be unified.
>
> [Brook(2004), section 3.4]

My aim is to show that both kinds of unity are also preconditions of speaking a language, or, to put it differently, that being a speaker is sufficient for meeting the transcendental conditions of consciousness. The requirement of a single common subject, to which all experience is ascribed, is entailed in being a speaker thanks to the demand for there to be *someone* taking responsibility for judgements, as explained in section 4.3.3. The second kind of unity is more complicated and Brook defines its sense as follows:

> The unity of consciousness $=_{df.}$ (i) a single act of consciousness, which (ii) makes one conscious of a number of representations and/or objects of representation in such a way that to be conscious by having any members of this group is also to be conscious by having others in the group and of at least some of them as a group.[Brook(2004), section 3.4]

The concept of unified consciousness is to account for the transcendental feature of our experience that we are conscious of a manifold in a single state of consciousness, and of having *distinct* conscious experiences in succession.[112] The consciousness of a manifold is, at least in one narrow sense, entailed by being a speaker in that to understand meaning of a claim, to employ concepts in a judgement, is to know what are the reasons for it and what other claims can be inferred from it. The manifold is united, so to say, thanks to the semantic relations that obtain among the contents of thoughts and which ought to be apprehended if the subject is to count as a speaker of meaningful claims.

5.1 The language-consciousness relation

Before I elaborate on these suggestions, let me clarify my view of the relation between language and consciousness. On the empirical side, I gather that linguistic capacity is constitutive of consciousness in that becoming able to speak *meaningfully* is the very same process by which one becomes self-conscious in the sense that comprises awareness, personhood and reflexion . On the logical side, it may be argued that consciousness as a *theoretical* concept (or entity) owes its criteria of correct application to (or manifests itself in) a pattern of linguistic expressions.

[112]This claim ultimately appeals to our intuition as it cannot be further argued for, given that we don't know what the identity of an act of consciousness consists in. Since the fact is transcendental, we may read it as saying that whatever *model* of conscious experience we entertain (e.g. the higher-order theory), it must explain how being conscious of something entails being conscious of something else.

There is a variety of kinds of empirical evidence that support the idea of the constitutive role of language for consciousness. For example, evolutionary stories and arguments can be identified as one such kind. Besides the general idea that human sapience coevolved with language, the evolutionary paradigm provides the perspective of purpose: for what benefit and under what selective pressures has consciousness evolved? This perspective may give us clearer idea what consciousness amounts to in natural terms. Another kind of evidence draws on consequences of some neurological impairment, such as lesions or brain injuries. There are various cases in which local brain damage causes aphasia (a kind of impairment of the linguistic capacity) that support the modular account of the linguistic capacity, for the afflicted aspect of the capacity is localized to the damaged part of the brain and is considered relatively independent of other aspects. Some aphatics, especially those who suffer from impairment in the reception of language, exhibit disruptions in their consciousness, such as retrospective confabulation, in which the subject reports she is not conscious of a stimulus and offers an unlikely but coherent interpretation of her own past behaviour, which betrayed some sort of awareness of the stimulus. Most notably, split-brain patients, whose hemispheres are surgically detached, insist that they are not conscious of what lies on the left side of their visual field while their behaviour (e.g. grasping things manually) shows that the information from the left visual field is available to other capacities.[113] Yet another kind of evidence may come from psychological studies, ranging from developmental psychology to psychoanalysis.

No doubt, putting the empirical evidence into the broader philosophical context is a subtle task that requires a great deal of knowledge in both philosophy and the empirical science concerned. We may settle for the view that empirical consciousness is causally related to linguistic capacity and take that as a reason for looking for similarities between language and consciousness at the transcendental level of their preconditions.

5.2 The speaking subject

I have followed Brandom's interpretation that the transcendental selfconsciousness is best understood in terms of relating (and so uniting) all commitments one has made by her judgements to a single subject who is responsible for them. Regarding any sentence as an expression of a judgement is *ipso facto* thinking of a subject that endorses it and is thus responsible for demonstrating its truth. When reading novels, we can clearly distinguish

[113] See [Nagel(1971)].

between fictional subjects that are committed to different claims: the narrator, various characters, the author. This is to show that the presupposition of a responsible subject is inherent to the concept of judgement no matter whether there actually *is* a physical embodiment of the subject. The concept of judgement also seems to presuppose there is a community whose members can hold each other responsible for their claims (and again, the community may be fictional as well). But taking responsibility for one's claims is a social practice that can be learned only from a censorious community using language. In consequence, wolf-children, or generally people that have never acquired the ability to communicate judgements in one way or another, are not self-conscious.[114] For those reluctant to accept such a consequence, it could be the reason to reject this language-oriented interpretation of transcendental self-consciousness; for my part, I agree.

I believe the preceding interpretation conforms to points 2 and 3 mentioned in 2.2.1 (p. 22). The subject presupposed by the act of endorsing a claim is not specified by anything else beyond the act itself, taken abstractly. The fact that in empirical consciousness one identifies the subject with one's body follows from the way the social practice of giving and asking for reasons is realized: in communication, we address each other's bodies as these are the things that express the judgement.[115] In the act of synthesis I am conscious only of that I am insofar I only have to "posit" a subject that endorses the claim. The same point is made by Cartesian *cogito*, the difference being only that where Descartes inferred that the subject is necessarily a thinking thing, no such specification is appropriate along the transcendental lines.

Yet a crucial question remains unanswered: what makes all the judgements being united in a *single* subject? From what has been said, it only follows that each judgement presupposes a subject that takes responsibility

[114]Thus when reading about Robinson Crusoe, we believe Robinson *thinks*, makes judgements, because we understand he has acquired this social ability before he shipwrecked on an island. It would be much less credible, I think, when reading a story about a wolf child. Significantly, in such stories (e.g. The Jungle Book or Tarzan) the animals are either depicted as very social or the character exhibits very simple stream of thoughts.

[115]I don't intend to imply that our selves are radically different from our bodies. I only want to emphasize that the identity of the transcendental subject is not constrained by any specific physical realization (it is, after all, a "logical entity") though it might necessarily be bound to some physical realization for the purpose of communication. Considering once again a talking computer, if its CPU, memory and other functional components were far away from the peripheries (keyboard, screen, robotic hands etc.), it is likely that the subject would be localized (by someone adopting the intentional stance) to the place where it *acts*, i.e. to the peripheries. Nevertheless, the transcendental subject would not "change" were the peripheries attached to the functional machinery. Detailed elaboration of the point can be found in [Dennett(1978)].

for it, not that the judgements have to be ascribed to one such subject. The general answer Strawson suggests, namely that it is thanks to our fundamental consciousness of exercising the power of synthesis, i.e. of judging, does not itself explain much. Consider, however, that in making a judgement one commits herself to other claims and justifies the judgement by reasons which again has to be claims she endorses. I would not endorse the claim "This is a dog." if I thought someone else was thereby committed to the claim "This is a mammal." If I am to mean it, I have to unite all logically related judgements to a single subject. Yet again one could object: there still could be isolated, coherent, self-contained sets of judgements such that endorsing a claim from one such set is independent of endorsing any claim from other sets; and how can judgements from different sets be united in a single subject? The only available answer is then that one is conscious not only of logical, normative relations but also of individual associations. For example, I know that something of the thought A made me think B, though there is no semantic relation between them and I may not be able to make the association link explicit. Admitedly, however, appealing to a kind of implicit knowledge of transition from one thought to another is once again relegating the explanation of unity of transcendental self-consciousness to some intuitive power that has no clear connection to linguistic capacity.

It is worth contrasting the preceding elaboration of transcendental self-consciousness with a similar idea which Dennett employs in explaining what the empirical self is.[116] He says self is like a fictional character constituted by the story of one's life; he calls the self a *center of narrative gravity*. Recalling his intentional stance, the claim may be rephrased as saying that the self is a unit to which intentions are ascribed, essentially an agent. The point is that our selves are no less fictional than selves of the others - we create our selves by the same process of intentional interpretation that we apply to others. Sometimes the behavioural pattern of an agent is so incongruous that the best interpretation is positing more selves - as in case of multiple personality disorder. The analogy with centers of gravity is to show that while selves are abstract, theoretic entities, they are determined by real things and can be employed in causal explanations. Unlike Kant, Dennett does not find it necessary to speak about transcendental self in the attempt to explain or describe human cognition. He thinks our cognition and conscious, mental life consists solely in the work of several subsystems whose "intentions" are united by a retrospective interpretation of my body as an agent. Thus Dennett believes only in the theoretical unity of an empirical self or the unity in the sense of availability to a single cognitive subsystem that may be regarded as crucial for consciousness, e.g. decision-making or language production.

[116] See [Dennett(1992)].

5.3 Speaking of meaning

I am inclined to accept the higher-order theory of empirical consciousness: to be conscious of some content is to be able to form a higher-order belief about the content, such as "I think that S." This is in accord with the Kantian idea that being conscious of something amounts to conceptualizing it in the act of synthesis of intuition and understanding, for only something with a conceptual structure can be part of the higher-order belief. Whether the higher-order belief must be actually tokened, as a brain state with appropriate representation, or whether one can settle for *disposition* to cause tokening of such a higher-order belief, depends on specific ontological constraints we put on minds; I will not address these issues.[117] Now, what is crucial for consciousness in entertaining a higher-order belief is not merely that it is *about* the nested content, thus rendering it conscious, but more specifically, that the higher-order belief's explicit ascription to oneself makes it more likely that the subject thinks about what it means. That is, psychologically speaking, thoughts of other claims, that are semantically related to the original one, are more likely to be activated as the subject realizes the commitments following from endorsing the original thought.

If we leave the transcendental paradigm and enter the psychological one for a moment, we can argue that various contents acquire different degree of consciousness depending on the effects on the behaviour of the agent.[118] Where there is no overt behaviour to be related to conscious thought, as in the case of inner speech, we can still consider a language-specific inner articulation of a thought to be the goal the contents "compete for". Intuitively, the act of inner articulation is derivative from the act of endorsing a claim in a community, for the stream of inner thoughts ought to, in order to be of any use for us, meet the demand for coherence as well. That is to say, criteria of selection of some content to inner articulation are semantic, among else. What I choose to say to myself, as well as what I say to others, depends on meaning. Now, can we accept the preceding intuition and simultaneously avoid the commitment to a token of a preconceptual message that is somehow represented in the brain and wished to be articulated? We can, I think. The stream of thoughts can be generated by semantic considerations aimed to achieve some general discursive goal. For example, I may have just finished reading a novel that I find rather bad. In the attempt to make explicit why I consider it a bad novel, I start justifying my judgement by considering various features of which I know that generally make a novel bad, such as clichés, plain characters or badly structured plot. By

[117] I have partly postponed the discussion to appendix A.

[118] Cf. Dennett's idea of competition for behavioural control, see quotation on p. 52.

introspection, I may find that I go through all these features and ask myself again if I can justify the claim that they apply to the novel. In effect, by building up such a justification I am becoming more conscious of the original thought that the book is bad. Similarly, I can start at a premise and see what follows from it. Undoubtedly, semantic associations are not the only psychological mechanism of transition from one thought to another. But the criteria for the selection to articulation may be essentially semantic. This requires that the subject has mastered semantic discourse in which one can argue for the correct use of an expression. As Foucault famously pointed out, the correctness is relative to a discourse, thus the same linguistic expression may entail different commitments in different discourses. The subject may well be aware of these differences and be able to switch from one discourse to another.[119]

Now, returning to the transcendental level, our thinking is autonomous in the sense that we choose what judgement we endorse. Our thinking is not a simple chain of "blind" associations, but rather a reasoning bound by the same norms that apply to public claims. Brandom emphasises in his discussion of Kant that if a subject is to choose what to endorse, what commitment to make, it presupposes that the subject takes the binding norms as independent of herself. Specifically, taking responsibility for the judgement "This is a dog." presupposes that I employ the concept DOG in accord with the norms which govern public usage of "dog".[120] Hence the subject must be able to make explicit the commitments pertaining to a judgement in order to decide whether to endorse it or not. This requires the ability of forming higher-order beliefs expressing semantic relations. Without such an ability (and its exercise) we could not recognize our thinking as free. For what other property than meaning could we decide to make a particular judgement? And if we could not *deliberate* on meaning and yet our thoughts were associated on the meaning's basis, would we not be under the impression that "something thinks for us" (if we ever stumbled upon this thought thanks to the semantic engine working in us)? Even though our thinking may be causally determined by brain processes at the *empirical* level, at the *practical* level we must regard ourselves and each other as free in what we say or think, since that is already implied by the concept of normativity. It seems to be a necessary feature of our conscious thinking that it regards itself as free, which again is a Kantian theme. If that is right, we have a good reason to

[119]While Foucault sticks to the relativism, others, like Brandom, I think, would argue that there must be a master discourse in which the implicit internal relations of a discourse can be made explicit to the benefit that the subject can orient herself in it. Naturally, the master discourse would be logic.

[120]See [Brandom(2006), section 9].

believe that the ability to think about meaning (itself derivative from talking about meaning) is constitutive of consciousness, for it underlies our choices of what judgement to endorse, which, to repeat, is the goal for attaining of which various candidates for conscious thoughts compete.

6 Conclusion

I have argued that the Hard problem consist in explaining transcendental consciousness: how is the manifold of experience united to a single subject so that the "I think" can possibly accompany any representation. Following Kant's hint that the uniting power is inherent to the synthesis whereby intuition are brought under concepts, and arguing that the conceptual is tightly linked to the linguistic, I have proposed to look in the domain of philosophy of language for ideas that could possibly contribute to our understanding of what constitutes consciousness. The discussion of the nature of meaning, intentionality, and speaking has indicated that there are striking similarities between the characteristic features of consciousness and the features recognized in language and its use.

The similarities support the idea that language is constitutive of consciousness. Here is the short explanation of the specific sense that the idea has acquired in this thesis. Mastering language requires that the speaker develops the capacity to think about meaning, to deliberate on what to say. This capacity is very complex and the conditions it entails have direct significance for consciousness. One ought to be able to recognize herself as a subject that takes responsibility for her judgements, and also to bear in mind the complex of semantic relations of a single judgements to many other judgements, which requires the ability to make the semantic relations explicit in a higher-order belief. Roughly speaking, the preconditions of consciousness, revealed by transcendental considerations, are fulfilled by the full-fledged linguistic capacity. All this converges to the idea that by "minding the language", one becomes conscious - and thus able to speak her mind.

A Carruthers's theory of consciousness

P. Carruthers in his (1996) monography puts forward a theory of consciousness that is supposed to combine the best of Fodor, Dennett and various findings from cognitive science. In a nutshell, his theory 1) regards consciousness in terms of availability for higher-order thoughts which again are reflexively available to higher-order thoughts; and 2) claims that natural language is *necessarily*[121] involved, at least in case of human beings, in constituting this higher-order availability structuring.

Let me mention most of the categories and -isms of philosophy of mind and language that Carruthers willingly classifies his theory into. First, he endorses the cognitive conception of language according to which language is not just a mean of communication but structures our cognition because we *think in* it. Second, he agrees with Fodor that thoughts, beliefs or propositional attitudes in general are to be conceived as relations to internal sentences. Unlike Fodor, however, he does not take it necessary to stipulate a language of thought for these internal sentences and contends that they are natural language sentences.[122] Third, he is a Fodor-style realist about mental states: tokens of mental states are tokens of some inner states of physical organization endowed with causal powers to bring about other mental states; the pattern of mental state relations, which is deduced by our folk-psychological reasoning, mimics a pattern of causal relations. See section 4.2.2 here or [Carruthers(1996), 1.6]. Fourth, he believes that the theory of mind behind our folk-psychological reasoning is largely innate, claiming that it is the best explanation of striking data from developmental psychology about how soon and reliably can young children operate on such a theory.

Carruthers makes it clear that he designs his theory to account for the property of consciousness of *mental states*, not of self-consciousness or phenomenal consciousness (having states with certain feel to them). However, he seems to accept the what-it-is-likeness parlance of Nagel and other proponents of qualia. Still, he employs the concept of phenomenal *feel* in a way which suggests that my identification of qualia, in their least controversial interpretation, with Kantian intuitions is correct. He starts at a thesis he

[121] Carruthers specifies the necessity as natural, as opposed to conceptual or metaphysical necessity. In effect, the constitutive character of language for human consciousness is an empirical matter, not a conceptual one, since it follows from our physical nature, environment and actual physical laws that obtain. See [Carruthers(1996), sections 1.1, 1.4].

[122] "When a speaker utters a sentence, on this view, their utterance expresses a thought by *constituting* it, not by encoding or signalling it. A hearer who is a competent user of the same language will then understand that utterance in virtue of it constitutively expressing, for them, the very same (...) thought."[Carruthers(1996), p. 2].

deems uncontroversial: "[F]or there to be conscious experience there must be something that the experience is *like*."[123] Next, he observes that *knowing* what it is like to be in some state entails being aware of having that state. Given that the subject may know what it is like for her to be in a state, she must be able to recognize and distinguish between its experiences as such; that is, not only the subject can respond differentially to blueberries and redberries, but she must be able to conceive of the difference between perceptions of blueberries and redberries. Hence, Carruthers infers the necessity of higher-order thought theory *from* the starting thesis about phenomenal feel of conscious experience. This seems to be a strange line of argument. As I have argued in 2.1.1, I don't think we can make a good sense of the starting thesis. Either we have to read "there must be something that the experience is like" as a strong ontological commitment, which runs into familiar difficulties, or it is too vague a claim to start an argument with.

Later in the book, however, Carruthers takes on different line, much more plausible in my opinion: "[P]henomenal feeling will emerge, of natural (perhaps metaphysical) necessity, in any system where perceptual information is made available to thought in analogue form, and where the system is capable of recognising its own perceptual states, as well as the states of the world perceived."[124] He subsequently notes that not only there are purely recognitional, non-inferentially applied concepts such as RED or LOUD, but the awareness of the fact that one has experience *as of red* is non-inferential as well.[125] Interestingly, he adds: "There might be a natural tendency to 'carve off' these recognitional concepts from their surrounding beliefs, and to use them independently - especially if we find the properties which they pick out to be of intrinsic interest to us."[126] In Kantian terms, this could be interpreted as an attempt to strip a perceptual experience of its conceptualization to get a peek at raw intuitions. But that is *conceptually* impossible; intuitions cannot be experienced, they are conceptualized. However, if the craving for experience of raw intuitions is diagnosed as a peculiar philosophical perversion originating from empirism, we at least know what the fallacy consists in: in the belief that the recognitional aspect of a concept has some reality independent of the concept's functional and causal role. "[T]he property of being the subjective feel of an experience of red is a functional one,

[123][Carruthers(1996), p. 154].

[124][Carruthers(1996), p. 212].

[125]Theoretically speaking, a system with suitable cognitive architecture (i.e. satisfying his reflection theory of consciousness yet to be described) "will have the capacity to classify informational states according to the manner in which they carry their information, not by inference (...) or description, but immediately.[Carruthers(1996), p. 213].

[126]ibid.

identical with possession of a distinctive causal role (the causal role namely, of being a state whose normal cause is red, and which is present to a reflexive thinking faculty with the power to recognise its own perceptual states as such).[127] Consequently, Carruthers as well as Dennett conclude that the well-known inverted spectrum thought experiment is idle; my sympathies lie with them.

In expounding his reflexive theory of consciousness (RT, henceforth), Carruthers employs modules and boxes in the same way as Fodor. Recall that for Fodor to believe that P is to have P in one's belief box, which is just different manner of expressing that the subject bears relation of believing to a sentence in LOT. Similarly, Carruthers says that being conscious of something is having that information in one's 'conscious box', or more specifically, having that information in a short-term memory store such that all information there is "available to acts of thinking which are reflexively available to further thinkings."[128] While schematized as a box, a part of short-term memory, Carruthers clearly conceives of its identity in functional terms; thus it is not an unexplained, stipulated black box, but rather a cognitive component that exhibits the quoted feature. It is his *sententialism* (see second and third characteristic of his theory at the beginning) which implies that the component must be conceived of as a special container for sentences. We can, however, refrain from endorsing any specific stance like sententialism, realism about the mental etc., and consider only the functional role of the component.

I think the functional specification Carruthers offers points in the right direction, so let me elaborate on it. He classifies his RT as a species of that gender[129] of higher-order thoughts theory of consciousness which states:

> Any mental state M, of mine, is conscious = M (level 1) is disposed to cause an activated belief that I have M (level 2), which in turn is disposed to cause the belief that I have such a belief (level 3), and so on; and every state in this series, of level n, is disposed to cause a higher-order belief of level $n + 1$. [Carruthers(1996), p. 174]

[127] [Carruthers(1996), p. 214].

[128] [Carruthers(1996), p. 194].

[129] Carruthers recognizes four alternative higher-order theories based on variations along two dimensions. The first dimension concerns whether higher-order beliefs must be actually physically tokened in order to be conscious of the subject matter of the belief or whether there just ought to be disposition for their tokening. The second dimension concerns the question whether higher-order beliefs are themselves conscious. His RT theory belongs to the alternative which lets the mental state be only disposed to cause a higher-order belief and requires them to be conscious too.

Crucially, the recursive account of consciousness does not entail infinite regress because what matters is the *disposition* of a state to cause tokening of a belief that I am in such state. Furthermore, the higher-order belief must itself be so disposed; thus it does not suffice merely to have a mental state that is *about* another (hence lower-order) mental state - unless the former is disposed to be reflected on. It is obvious why language fits nicely in this scheme, for the recursive character is attained by a simple syntactical operation of prefixing the original sentence with "I think". But the trivial 'logical' disposition of sentences qua linguistic abstractions (i.e. not as mental tokens) for such *syntactical* extension does not ensure that every sentential mental state will be disposed to *cause* some higher-order state. Only the right ones will be, the conscious ones; we do, after all, experience strange moments of articulating some meaningful linguistic expression (mostly in a kind of inner speech or quiet muttering) without being aware of it. So linguistic expression is not sufficient for consciousness, according to Carruthers. It is, however, naturally necessary for it in the following sense:

> Some human conscious thinking is such that, of natural necessity, it involves public language (in virtue of the given architecture of human cognition, together with causal laws); and, necessarily, some of these propositional thoughts belong to types which (for us at least) constitutively involve such language.[Carruthers(1996), p. 263]

To complete the picture, Carruthers has it that the folk-psychological module of our cognition enables us to conceptualize our occurent thinkings, although we are aware of what *we* think non-inferentially, unlike in case of other minds. In other words, the classificatory concepts like THOUGHT, BELIEF or DESIRE come from the cognitive module responsible for our implicit theory of mind.[130] One has to possess these concepts in order to reflect on her own thinkings, i.e. to be aware of her own thoughts, beliefs, etc.

[130]That it is a relatively independent module and not an acquired and culture-dependent theory, like "folk literary theory", is presumably substantiated by early signs of the theory working in young children's behaviour, much earlier that they master the language in a way sufficient for learning the theory explicitly. Admittedly, the evidence from developmental psychology may not be as conclusive as interpreted above, but still we can at least agree that we are born predisposed to master folk-psychological reasoning. Anyway, it does not follow that the folk-psychological concepts ought to be ready for use in minds of wolf-children - probably being exposed to social interaction is necessary for the concepts to get established in one's mind. On the other hand, it seems to follow that the concepts would be in force in cognition of someone who does not speak any language, yet who has not been deprived of social interaction and whose mental condition is not severely impaired (if such a situation is conceivable at all).

Interestingly, as Carruthers needs to allow that both conscious and unconscious thinking involves natural language, he resorts to imply that it is *phonological* articulation (presumably in one's auditory imagination) what grants access to thought's content by higher-order thoughts.

> My hypothesis is that it is by formulation some of our occurrent thinkings in the form of images of natural language-sentences that our cognitive system is able to gain access to (in such a way as to render conscious) its own processes of thought. The function of the phonological loop is thus much more than just to enable the system to engage in language-involving processing tasks. It is also to enable the system to gain access to its own occurrent thoughts, thus facilitating the sort of indefinite self-improvement that comes with self-awareness. And according to RT theory the sentences represented in the phonological loop often *are* the acts of thikning the thoughts wihch are expressed by those sentences.[Carruthers(1996), p. 247][131]

In reply to a hypothetical objection from advocates of LOT that any thought in natural language might be *ipso facto* conscious and that thinking as well as language production presuppose merely *conceptual* (not yet articulated) level of representation, Carruthers appeals to purely semantic level of language expressions - like lemma level of an entity of mental lexicon in psycholinguistics, or logical form in Chomskian framework.[132] One unconcsciously thinks, as it were, at this semantic level free of vocalization, that nevertheless is bound to a specific language. This opens a huge debate as to what exactly this semantic or conceptual level is supposed to be. Although it is conceivable that experiments may be designed such that would test the hypothesis that the conceptual is (native) language-specific, it nevertheless blurs the dividing line between LOT and RT theory of consciousness, since both operate with *representation* in a medium that is semantically evaluable and to a great extent abstract. I am actually inclined to believe that some such level of representation has to be *inevitably* presupposed if we attempt to naturalize consciousness. Likewise I believe that *meaning* of such representations ought to be derivative from meaning of their articulated expressions.

[131]The phonological loop is a part of 'working memory' model presented first by A. Baddeley a G. Hitch in their 1974 article 'Working Memory', in *The Psychology of Learning and Motivation*, vol. 8, ed. G. Brown, Academic Press.

[132]Cf. [Carruthers(1996), pp. 249,267].

References

[Brandom(2001)] Robert Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, 2001.

[Brandom(2006)] Robert Brandom. Kantian Lessons about Mind, Meaning, and Rationality. *The Southern Journal of Philosophy*, 1 January 2006.

[Brook(2004)] Andrew Brook. "Kant's view of the mind and consciousness of self", *The Stanford Encyclopedia of Philosophy*, 2004. URL `http://plato.stanford.edu/entries/functionalism/`. This is an electronic document. Date of publication: July 26, 2004. Date retrieved: October 28, 2009. Date last modified: October 20, 2008.

[Carruthers(1996)] Peter Carruthers. *Language, thought and consciousness: an essay in philosophical psychology*. Cambridge University Press, 1996. ISBN 0-521-63999-9.

[Chalmers(1995)] David Chalmers. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, (2):200–19, 1995. URL `http://consc.net/papers/facing.html`.

[Chalmers(2003)] David Chalmers. Consciousness and its Place in Nature. In S. P. Stich and T. E. Warfield, editors, *Blackwell Guide to the Philosophy of Mind*, pages 102–142. Blackwell, Oxford, 2003.

[Dennett(1978)] Daniel C. Dennett. Where Am I? In *Brainstorms*. Bradford Books, 1978.

[Dennett(1992)] Daniel C. Dennett. The self as the center of narrative gravity. In *Self and Consciousness: Multiple Perspectives*. Lawrence Erlbaum, 1992.

[Dennett(1988)] Daniel C. Dennett. Quining Qualia. In A. Marcel and E. Bisiach, editors, *Consciousness in Modern Science*. Oxford University Press, 1988.

[Dennett(1991)] Daniel C. Dennett. *Consciousness Explained*. Little Brown, New York, 1991.

[Dennett(1996)] Daniel C. Dennett. *Kinds of Minds: The Origins of Consciousness*. Phoenix, 2001. ISBN 0-75380-043-8. First published in 1996 by Weidenfeld and Nicolson.

[Dennett(2006)] Daniel C. Dennett. The Evolution of "Why?": Essay on Robert Brandoms *Making it Explicit*. (unpublished) 2006.

[Fodor(1987)] Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. The MIT Press, 1987.

[Fodor(2008)] Jerry A. Fodor. *LOT 2: The Language of Thought Revisited*. Oxford University Press, 2008.

[Haugeland(1990)] John Haugeland. Intentionality All-Stars. *Philosophical Perspectives*, 4:383–427, 1990.

[Kant(1781)] Immanuel Kant. *Critique of Pure Reason* (translated by P. Guyer and A. Wood). Cambridge University Press, 1997. ISBN 0521657296.

[McGinn(1989)] Colin McGinn. Can We Solve the Mind-Body Problem? *Mind*, 98(391):349–366, July 1989.

[Millikan(2005)] Ruth Garrett Millikan. On Meaning, Meaning, and Meaning. In *Language: A Biological Model*, pages 53–76. Oxford University Press, 2005.

[Nagel(1971)] Thomas Nagel. Brain bisection and the unity of consciousness. *Synthese*, 22(May):396–413, 1971.

[Nagel(1974)] Thomas Nagel. What Is It Like to Be a Bat? *Philosophical Review*, pages 435–450, 1974.

[Peregrin(2005)] Jaroslav Peregrin. Is Compositionality an Empirical Matter? In Markus Werning, Edouard Machery, and Gerhard Schurz, editors, *The Compositionality of Meaning and Content.*, volume 1. Ontos Verlag, 2005.

[Prinz and Clark(2004)] Jesse Prinz and Andy Clark. Putting Concepts to Work: Some Thoughts for the Twentyfirst Century. *Mind & Language*, 19(1):57–69, February 2004.

[Searle(1990)] John R. Searle. Is the brain's mind a computer program? *Scientific American*, 262(1):26–31, 1990.

[Sellars(1974)] Wilfrid Sellars. *Essays in Philosophy and its History*, chapter "... this I or he or it (the thing) which thinks...", pages 62–88. Springer, 1974.

[Sellars(1956)] Wilfrid S. Sellars. Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science*, 1, 1956.

[Shields(2000)] Christopher Shields. "Aristotle's psychology", *The Stanford Encyclopedia of Philosophy*, 2000. URL `http://plato.stanford.edu/entries/aristotle-psychology/`. This is an electronic document. Date of publication: January 11, 2000. Date retrieved: October 28, 2009. Date last modified: April 28, 2003.

[Strawson(1966)] Peter Frederick Strawson. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. Routledge, London, 1966. ISBN 978-0415040303.

[Szabó(2004)] Zoltan Gendler Szabó. "Compositionality", *The Stanford Encyclopedia of Philosophy*, 2004. URL `http://plato.stanford.edu/entries/functionalism/`. This is an electronic document. Date of publication: April 8, 2004. Date retrieved: October 28, 2009. Date last modified: February 14, 2007.

[Turing(1950)] Alan Turing. Computing Machinery and Intelligence. *Mind*, (59):433–460, 1950.

[Wittgenstein(1953)] Ludwig Wittgenstein. *Philosophical Investigations* (translated by G. E. M. Anscombe. Blackwell, second edition, 1958. ISBN 0-631-20569-1.