FILOZOFICKÁ FAKULTA
Univerzita Karlova

DISERTAČNÍ PRÁCE

Mgr. Ing. Martin Vraný

Naturalizing the Unity of Consciousness:

can neuroscience explain a fundamental feature of subjectivity?

Naturalizace jednoty vědomí:

mohou neurovědy vysvětlit zásadní rys subjektivity?

Ústav filosofie a religionistiky

Vedoucí disertační práce: prof. RNDr. Jaroslav Peregrin, CSc.
Studijní program: Filozofie
Studijní obor: Filozofie

Praha 2017

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

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

I would like to thank my supervisor Prof. Jaroslav Peregrin for his patience with the slow progress of this work. I am very grateful to The Fulbright Commission for the scholarship that allowed me to spend a most inspiring year at UC Berkeley and Prof. John Searle for the invitation. I am also grateful to The Anglo-Czech Educational Fund for the funding of my stay at The University of Edinburgh and to Prof. Andy Clark for his supervision of the part related to the predictive coding theory.

Abstrakt

Cílem disertační práce je analyzovat pojem jednoty vědomí coby explanandum pro přírodní vědy a zhodnotit, do jaké míry se daří hlavním neurovědeckým teoriím vědomí tuto jednotu vysvětlit. Práce je motivována přesvědčením, že právě jednota vědomí představuje největší výzvu, které čelí vědecké pokusy vysvětlit vědomí.

V úvodu práce tvrdím, že důvodem proč některé teorie vědomí vedou k tomu, co Dennett nazývá karteziánským materialismem, je právě to, že se dostatečně nevěnují problému jednoty vědomí. Pokud bychom dobře rozuměli jednotě vědomí a její roli v přírodě, snáze bychom se vyhnuli tendenci vykládat vědomí způsobem, který je vskrytu homunkulární.

Druhá kapitola analyzuje různé aspekty, z jejichž hlediska se vědomí považuje za jednotné. Dva z těchto aspektů představují obzvlášť velkou výzvu při naturalizaci vědomí a zároveň jsou spolu neslučitelně spjaté. Jde o jednotu vědomých obsahů v určitém čase a jednotu ve smyslu jednoho subjektu majícího vědomé obsahy, schopného na ně reflektovat.

Třetí kapitola se zabývá hlavními pojmovými a metodologickými obtížemi, které stojí před každým naturalistickým výkladem jednoty vědomí. V kapitole rozlišuji normativní a objektivní aspekty pojmu jednoty vědomí a ukazuji, že vysvětlit objektivní aspekt jednoty je úkol pro kognitivní neurovědu. Zároveň zde obhajuji pojem neuronální reprezentace, bez kterého by naturalizace jednoty byla nemyslitelná.

Ve čtvrté kapitole se věnuji Kantově transcendentální psychologii za účelem detailnější analýzy pojmu jednoty vědomí. Kant podal velmi promyšlený výklad jednoty vědomí pod názvem transcendentální jednota apercepce a zavedl důležité rozlišení mezi empirickým a transcendentální sebe-vědomím. Kantovy argumenty tak zasazují jednotu vědomí do širšího kontextu poznávání světa, situovanosti a aktivního působení.

V paté kapitole představuji filosofické úvahy, které pomáhají specifikovat jednotu coby explanandum pro přírodní vědy. Nejprve se věnuji práci Shoemakera a Castenedy o logice referování k sobě samému. Následně představuji Hurleyové model vědomí založený na tzv. dvouúrovňové závislosti a též její pojem perspektivního sebe-vědomí. Její model zachycuje jednotu vědomí v celé šíři jejího pojmového rozsahu a jeho formulace je zároveň v souladu s řadou teorií z kognitivní vědy. Na závěr představuji výklad empirického sebe-vědomí, který spojuje řadu bodů představených v předchozích kapitolách. Hlavním tvrzením v této části je, že stav sebereflexe bychom měli chápat jako reprezentační transformaci reflektovaného stavu, nikoli jako stav vyššího řádu, který by reflektovaný stav obsahoval jako vlastní část.

V závěrečné šesté kapitole diskutuji vybrané neurovědecké teorie vědomí: teorie neuronálního globálního pracovního prostoru, teorii prediktivního kódování a teorii integrace informací. Ukazuji, v čem jednota vědomí podle těchto teorií spočívá, a následně hodnotím, jak dobré je dané vysvětlení a jaké jsou jeho meze. Celkově docházím k závěru, že nejlepší, byť neúplné, vysvětlení poskytuje teorie prediktivního kódování.

Klíčová slova: jednota vědomí, neurovědy, naturalizace, filosofie mysli

Abstract

The aim of the dissertation is to analyze the concept of the unity of consciousness as an explanandum for natural sciences and assess how good an explanation do leading neuroscientific theories of consciousness provide. The motivation behind this project is the idea that it is the unity which poses the greatest challenge for the scientific quest for consciousness.

I argue in the Introduction that the reason why some theories of consciousness lead to what Dennett calls Cartesian materialism is precisely because they fail to address the problem of the unity of consciousness. If we had a good understanding of the unity of consciousness and its place in nature, we could more easily avoid the tendency to devise accounts of consciousness that are homuncular in disguise.

In chapter 2 I analyze various aspects in which consciousness is thought be unified and conclude that two such aspects are particularly challenging for naturalizing the unity and that they cannot be treated separately. They are the unity of conscious contents at a time and the unity in the sense of a single subject having conscious contents and being able to reflect on them.

Chapter 3 describes main conceptual and methodological issues faced by naturalistic accounts of the unity. I distinguish between a normative and an objective aspect of the unity and show that explaining the latter is a domain of cognitive neuroscience. I also defend the concept of neural representation without which naturalization of the unity is inconceivable.

In chapter 4 I turn to Kant's transcendental psychology in order to analyze the unity further. Kant provided a very detailed and insightful account of the unity under the term transcendental unity of apperception and drew an important distinction between empirical and transcendental self-consciousness. Kant's arguments put the unity of consciousness into the broader context of cognition, situatedness and agency.

In chapter 5 I present philosophical accounts that help to specify the unity as an explanandum for natural sciences. First, I review Shoemaker's and Castaneda's work on the logic of self-reference. Second, I present Hurley's two-level interdependence model of consciousness and her concept of perspectival self-consciousness. Her account not only preserves the conceptual richness of the unity as presented earlier but is congenial to many theories in cognitive science. Finally, I present an account of empirical self-consciousness that puts together the threads explored in previous chapters. The main point is to argue that the self-reflective state should be understood as a representational transformation of the object state, not as a higher-order state that contains the object state as its proper part.

In chapter 6 I finally review selected neuroscientific theories of consciousness: the neural global workspace theory, the predictive coding theory, and the information integration theory. I specify what the unity amounts to in these theories and assess how good their explanation is. Overall, I conclude that the predictive coding theory offers the best, albeit incomplete, explanation.

Keywords: unity of consciousness, neuroscience, naturalization, philosophy of mind

Contents

1 Introduction

Scientific study of the human mind, once thought to be impossible,[1] is now thriving. Since the cognitive revolution and the retreat of behaviorism, empirical research of mental processes and structures has been considered a legitimate pursuit; and thanks to new experimental methods quite fruitful as well. The invention of new brain-imaging devices, covering previously unreachable niches of temporal and spatial resolution, inspired optimism that we would soon have a causal explanation of the mind. Philosophers have often cautioned against putting too much hope to the neuroscientific research; and scientists, undeterred by philosophers' conceptual arguments, continued extending the knowledge of how the brain works in the piecemeal fashion typical of life sciences.

The more complex psychological phenomena were studied, the clearer it was that consciousness poses a great methodological and conceptual problem. It eventually became clear that the study of central cognitive processes, like attention or decision-making, could not be postponed indefinitely. Insofar as consciousness is implicated in those processes, a psychological account of consciousness was needed. Chalmers (1995) famously made a distinction between easy problems of consciousness and the hard problem, defined as the subjective aspect of experience: explaining what it is like to experience something using a third-personal objective description. His article set a long debate among philosophers opposing reductionism, scientifically inclined philosophers, and philosophically inclined scientists about the prospects of the science of consciousness. As a consequence, the philosophical debate shifted towards the (pseudo)problem of qualia to the

[1] Cf. Kant's *Metaphysical Foundations of Natural Science* where he argues for the impossibility of psychology as a proper science.

detriment of discussing the easy problems[2] whose philosophical reflection would still be very helpful to the empirical research and which are not so easy after all, as Chalmers himself admits.

I think that Chalmers's agenda had misguided philosophical efforts for some time and that a lot of useful philosophical work is there to be done on the "easy" problems of consciousness. Specifically, if we take seriously the possiblity that consciousness might be a matter of complex interaction of mental capacities which can be studied, to some extent, in isolation, the study of the interaction itself will become paramount for the prospect of explaining consciousness. A convincing account of the interaction would also help us to escape the grips of a false picture of the mind which Dennett calls Cartesian materialism; for now the lack of such account leads cognitive scientists to defer crucial parts of their explanations of the easy problems to not yet available accounts of the central processes - a promise that may never be fulfilled. The lack of such account, and with it the tacit operation under the false picture, impedes the progress in scientific explanation of the mind. In my understanding, explaining the unity of consciousness is a necessary step for moving forward in the scientific explanation of the mind and to ease the grip of Cartesian materialism.

1.1 Cartesian materialism debunked

Dennett (1991) goes a long way trying to expose as untenable the position of Cartesian materialism - the idea that there is a place in the brain where all contents come together and the subject thereby becomes conscious of them. Cartesian materialism is a remnant of the Cartesian Theater view inherent in dualism according to which the subject becomes conscious of something by "seeing" it presented in a place where the body and the mind come together. Even when dualism is openly discarded, this obviously problematic picture is often tacitly

[2]According to the original formulation, these are: "the ability to discriminate, categorize, and react to environmental stimuli; the integration of information by a cognitive system; the reportability of mental states; the ability of a system to access its own internal states; the focus of attention; the deliberate control of behavior; the difference between wakefulness and sleep." (Chalmers, 1995, p. 201)

substituted by the no less problematic picture of a material place or a process responsible for conscious seeming (over and above the process of discriminating the represented content itself).

Dennett identifies various sources of this mistaken view. One of them is the divide and conquer strategy of researching various psychological phenomena by focusing on either their input side or output side, and limiting the scope of subject's conscious decision-making to the minimum by careful experimental design. As a result, we have a good understanding of the processes thought to be peripheral, such as sensory processing or motor control, and a relatively poor understanding of the central processes such as attention, decision-making, or reasoning.[3] The explicit classification of the easy problems of cognitive science as the study of peripheral processes and the hard problem as the study of central processes perpetrates the sandwich picture of the mind as standing between input

[3]"Almost all researchers in cognitive science, ... , tend to postpone questions about consciousness by restricting their attention to the 'peripheral' and 'subordinate' systems of the mind/brain, which are deemed to feed and service some dimly imagined 'center' where 'conscious thought' and 'experience' take place. This tends to have the effect of leaving too much of the mind's work to be done 'in the center,' and this leads theorists to underestimate the 'amount of understanding' that must be accomplished by the relatively peripheral systems of the brain." (Dennett, 1991, p. 39) Although a great amount of research has been done since Dennett's analysis of the state of cognitive science 25 years ago, the situation is only slightly better. For example, Wegner (2005) repeats Dennett's attempt to debunk the homuncular idea of a controller (be it a process, a brain module, or some other functionally isolable thing), suggesting that the idea of a self exercising conscious control of behavior is a mere illusion.

One of the reasons for this, which is rarely mentioned, is that researchers opt for studying relatively peripheral cognitive capacities as independently of context as possible. The rationale behind this is that in order to establish a causal link between the dependent variables and the independent variables (which is the aim of any research hoping to provide a causal explanation), they need an experiment producing statistically significant results. Statistical significance is inversely related to the variance in data, and the variance increases with experimental conditions that are not controlled for and with the complexity of the investigated phenomenon. The more complex and interesting the psychological phenomenon, the more difficult it is to design an experiment that could yield conclusive evidence. In effect, if a complex psychological concept (e.g. attention) is eventually investigated experimentally, the concept is then operationalized in such a specific way that common sense finds the results only marginally relevant for understanding the (folk-psychological) concept that originally motivated the research.

and output that Hurley (1998) and Dennett (1991) show to be mistaken (or, at best, the proverbial ladder that should have been thrown away already).

Another important source for the mistaken view is the application of the appearance/reality distinction to the subjective domain itself: thinking that there is a fact of the matter about how things seem to someone, independent of what the subject *thinks* they seem to her. As Dennett puts it, it is natural to say "I judged it to be so, because that's the way it seemed to me." and understanding it rather literally as a description of a causal relation between two distinct events: the event of seeming to me (something is immediately given) and the event of me judging how things seem to be (my interpretation of the given). Dennett argues that there are no facts about the stream of consciousness independent of our ways of reporting it, one way or another. To explain why people nevertheless think otherwise (for that itself is a psychological fact that needs to be explained), he puts forward a theory of why some people might think that there are two distinct events and hence what forms the ground of the reality/appearance distinction for consciousness in their view:

> Some people presume that this intuition is supported by phenomenology. They are under the impression that they actually observe themselves judging things to be such *as a result of* those things seeming to them to be such. No one has ever observed any such thing "in their phenomenology" because such a fact about causation would be unobservable (as Hume noted long ago). (Dennett, 1991, p. 133)

Dennett's charitable explanation of the intuition would probably be that it is a product of folk-psychological reasoning which, while useful for everyday purposes of predicting other people's behavior, may be thoroughly mistaken, as he suggests at the beginning of his *Consciousness Explained* and *Intentional Stance*.

But Dennett is too quick dismissing that there be a substantial ground for the intuition. Indeed, the intuition is to some extent supported by phenomenology, namely by the experience of reflecting on one's current content of consciousness. To say the least, it makes sense to describe self-reflection as two distinct mental events in which the tokening of the object thought is constitutive of tokening of the reflective thought. Dennett might argue that both the object and the

reflective thought are judgements, so it supports only a rather innocent distinction between a lower-order and a higher-order thought, not a distinction between the apparent and real seeming. The object mental state is not the mythical given (Cf. Sellars (1956)). But this is easy to miss, since the reflective act is intuitively more judgement-like than the perceptual discrimination which might provide the content for the object thought.

The intuition is perpetrated by two facts. First, we can at will reflect on current contents of consciousness. Second, we can immediately realize that the subject of the reflective thought is the same as the subject of the object thought.[4] The first fact is intuitively explained by thinking of reflective consciousness on the model of perception of external objects: reflecting on one's occurent conscious contents is like perceiving objects in the world. Perception is a process by which an agent (an organism) becomes aware of an external object whose existence is independent of her. By analogy, conscious reflection is a process by which the self becomes aware of internal objects whose independence assumption (the real seeming, independent of our judgements about how things seem to us) is carried over from the perceptual model. How else could we explain this availability of conscious mental states for reflexion if not by positing them as independent, ready to be looked upon in a similar way that physical objects are always out there ready to be perceived? (This, to be clear, is a rhetorical question.)

It is one thing to recognize and argue against the homuncular regress inherent to the perceptual model of consciousness and self-consciousness, and another thing to offer a convincing picture that would replace the model so that it would lose its grip. One problem with Dennett's replacement, the Multiple-Drafts Model, is that it is still couched in mentalistic terms that presuppose a subject, or at least a point of view.

> These [processes of content creating, alteration, etc.] yield, over the course of time, something *rather like* a narrative stream or sequence, which can be thought of as subject to continual editing by many pro-

[4] I am not saying now that this identification is given in any other way but as a content of a higher-order thought whose object is the relation between the lower-order reflective state and its object. That is, the identity may be merely represented, and as such it could be wrong.

> cesses distributed around in the brain, and continuing indefinitely into the future. Contents arise, get revised, contribute to the interpretation of other contents or to the modulation of behavior which then eventually decay or get incorporated into or overwritten by later contents, wholly or in part. This skein of contents is only rather like a narrative because of its multiplicity; at any point in time there are multiple drafts of narrative fragments at various stages of editing in various places in the brain. (Dennett, 1991, p. 135)

The problem with this account is not the multiplicity of narratives, but the concept of narrative itself. Narrative for whom? I am not saying that Dennett does not have a good answer to this. I only want to point out that the metaphor of an observer in the perceptual model was replaced by a metaphor of a listener in the Mutliple Drafts model while asking us to imagine something like a free-floating narrative that creates its own listener:

> The Multiple Drafts model makes "writing it down" in memory criterial for consciousness; that is *what it is* for the "given" to be "taken" - to be taken one way rather than another. There is no reality of conscious experience independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory). (Dennett, 1991, 132)

Dennett here carefully avoids direct reference to a subject by using passive voice (to be "taken") but that does not help us understand how the subject can be dealt with. To explain the subject, Dennett employs a strategy typical for eliminativism - he argues that selves are useful fictions but not much more than that. The self is the center of narrative gravity, a theoretical posit that, just like the center of gravity in physics, is useful for predictions but whose existence is not independent of the things that justify its attribution.

Now, I agree that this line of thinking is relevant for the understanding of the personal, empirical self. But that is not the whole story, for there is also what philosophers call the transcendental self. The closest Dennett gets to discussing the transcendental self is when he considers that the sentence "This is *my* body."

does not mean the same thing as the sentence "This body owns itself."[5] And here he proceeds by asking what would it take for two or more selves to have the same body. He then goes on considering multiple personality disorder to further back his idea that the self can mean only a center of narrative gravity, interpreting MPD as an abnormal but perfectly intelligible case where two or more narratives are centered around the same body. What he misses is that the meaning of the sentence is not fully elucidated by considering what it would mean for a speaker to say that the speaker's body is not his or what could cause two different persons to claim ownership of the same body. A crucial part of the meaning, indeed the part which I would say is responsible for the reluctance of Dennett's readers to accept his theory of the self as complete, is that it involves what Shoemaker (1968) calls the subject use of the first-person pronoun 'I' (here guised in the possessive pronoun "my"), which is a marker of a specific kind of self-awareness.[6] In a nutshell, the subject use of 'I' presupposes a way of recognizing oneself that is not mediated by recognition of one's properties, because that would beg the question of how I know that I exemplify these properties. This non-inferential kind of self-awareness (henceforth transcendental self-awareness, see section 4) is what the second fact about consciousness (see above) is pointing at: we know, non-inferentially, the identity of the reflective subject and the subject of the object thought.

I think it is this aspect of consciousness that still sometimes lures us to the entrapment of some covertly homuncular view, despite all the progress that has been made in the consciousness science since the cognitive revolution. This aspect puts us on the horns of the following dilemma. On the one horn, we could try to explain away the transcendental self-awareness as mere part of the represented self-reflective content: it is a thought about the identity of two fictional objects - selves. This invites the question: for whom is this identity being represented? Perhaps we could try to turn this question on its head and reply that the fictional self is a by-product of such representation, not its creator (which would be the case if the perceptual model of self-consciousness were right). Dennett's theory and Rosenthal's higher order theory of consciousness follow this line of thought. But

[5](Dennett, 1991, pp. 418-422)

[6]See section 5.2 for a detailed account of Shoemaker's argument.

then the questions arise: how is the representation of the identity put together? One reason why it is natural to think that representation presupposes, rather than constitutes, a subject is that the posited subject functions as the principle of unity of the representation - the subject is what binds representations together into a complex one. To illustrate this simply: ten people in Ancient Greece could have entertained, at one moment, the contents that are part of the representation of Kepler's Second Law of planetary motion, but the representation of the law was not consciously tokened until Kepler "put them together", in one consciousness. Thus the unity of consciousness (the unity of represented contents) needs an explanation that goes beyond appealing to more content.[7]

The idea that it is the subject that puts the contents together is intuitive until we attempt to account for subjectivity and consciousness in material terms. This brings us to the second horn of the dilemma. We could say that there is such a thing (the subject) that perceives outer things through senses and its own mental states are transparent to it, so that it is immediately aware of them. Put like this, the view obviously commits the homunculus fallacy. That is why we usually find it in the disguised form which Dennett calls Cartesian materialism, where the subject is given a functional specification (e.g. as the place where or process whereby information converges and is transformed into a meaningful whole) that is supposedly physically realizable. Note that the Cartesian materialism picture with a functionally specified subject (e.g. as central processing) is problematic only relative to our explanatory project. If, for example, the task were to explain perceptual illusions, we could well use this picture and explain the illusions in terms of information processing taking place before the perceptual information reaches the subject.[8] Thus for some mind-related research projects this view could still be a useful simplification. Obviously, it won't do for the project of scientific explanation of consciousness.

If the dilemma indeed results from having a wrong picture about perception and cognition (one in which the subject receives information about the world,

[7]See Hurley (1998) and her just-more-content argument against such attempts to explain the unity of consciousness. The argument is recounted in chapter 3.

[8]This explains why several advances in cognitive science could have been made even though this picture is wrong as a global account of perception and cognition.

unites them, and decides on actions), the way out of the dilemma is to stop trying to explain how such a thing as the subject could or could not be realized, and focus instead on explaining those phenomena that invite this picture.[9] A naturalistic theory of consciousness can be right but unconvincing. For the theory to be acceptable, it should show where our folk-psychological picture of the mind is wrong and give a naturalistic account of those parts of folk-psychology that are right. What the phenomena inviting the Cartesian picture of the subject have in common is that they are all variations on the general notion that consciousness is unified.

1.2 Homuncular subject and the unity of consciousness

If we look at the subject as a folk-psychological concept, we may ask what is its explanatory role.

A prominent role of the subject as a folk-psychological concept is to guarantee personal identity, despite constantly changing body and contents of consciousness.[10] The sense of personal identity is in turn a prerequisite for social interactions and institutions. The subject is that what remains constant, essentially unchanging (though its attributes or affections are changing). Assuming identity for a thing is intuitively less problematic than understanding a sequence of events as a single process, but the challenges of accounting for the identity are

[9]This is Dennett's and other's strategy too, the difference is perhaps only in emphasising different aspects of consciousness as those that promote the wrong picture.

[10]It could be argued that the distinction between persons and subjects is a philosophical one - in the folk-psychological understanding it is conflated. Still, the popular knowledge of psychopathologies such as multiple personality disorder (which helps non-philosophers make the distinction between a person and a subject by illustrating that while there could be many persons exhibited by one body, there is always only one conscious point of view at a time. Furthermore, the intuitive appreciation of subjectivity manifests itself in considerations of solipsism, Matrix-like scenarios or even Descartes's *cogito*. The fact that people, often at young age, come up with similar ideas without previous exposure to Western philosophy suggests that the concept of subject is not (only) an academic invention but rather that it corresponds to some aspect implicit in our folk-psychological reasoning.

the same. When we face a complex phenomenon that we intuitively understand as one, there is nothing easier to cement its identity in the eyes of others than to give it a name and refer to it using the singular expression. Once the expression is in use, it is difficult to question the reasons that led us to thinking that the complexity is really a result of a single phenomenon or process. For example, using the word 'terrorism' to cover all attacks on civilians by military groups outside state armies leads us to assume that there is a single pattern (causal, sociological, or political) behind all those attacks - an assumption that might be wrong but which is rarely questioned because it is implicitly entailed in using the singular expresson. A related example from cognitive science is memory: researchers now widely agree that what folk-psychology recognizes as a single phenomenon is in fact an arbitrary assembly of cognitive processes that have very little, if anything, in common.[11]

Loosely speaking, the subject is in charge of doing various things (moving, talking, recalling, perceiving, thinking, seff-reflecting, creating personal narrative etc.) that appear to be coordinated and serving a common purpose. If we cannot explain the unity or coordination of the capacities that a conscious being exhibits, we can simply stipulate the unity be positing the subject and hope that one day the unity will be properly explained.[12]

The primary explanatory role of the subject is thus being that what binds together the many cognitive and conative faculties that an organism needs to succesfully navigate the world. But why stipulate the subject as conceptually distinct from the organism itself? Why could we not say that the faculties are unified simply in virtue of being realized by the same organism? The reason, I

[11]See Irvine (2012) for the discussion about the concept of memory and consciousness in science.

[12]It is important to recognize that this strategy is valid. Historically, people have recognized many patterns in nature for which they had no explanation. Naming the pattern helps to focus our thinking about the pattern and therefore to accumulate knowledge about it to the point when we understand why this pattern occurs, i.e. to the point when we explain the pattern at one level of description (say perceptual) by a regularity at another level (say biochemical). The old concept of life, in the sense of *élan vital* or some other unique property, had seemed to be irreducible to a physical description but a focused research eventually rendered that reduction possible.

think, is that it is not hard to imagine cases in which the coordination fails and the organism is still alive. States of unconsciousness, madness or infancy challenge the idea that all it takes to coordinate the capacities is to put them in one body. Without their coordination, the organism does not exhibit the pattern of interaction with the environment that we call being a conscious subject. The concept of person, crucial in social cognition and moral reasoning, further complicates this issue because it is more abstract than the concept of an organism. Not all bodies are persons and some bodies may have more than one. This promotes the notion that there is something with mental properties that is conceptually, if not ontologically, independent of the body. And as mentioned above, self-awareness and the contrast between the scope of self-knowledge and other-knowledge bring another twist to it by inviting us to think of self-reflection in terms of perception and some kind of transparency of the mind to itself.[13]

Thus to abandon the homuncular picture lurking behind the concept of a subject, we need to explain the unity of consciousness and then show again how it is related to the phenomena that are associated with it in the traditional view of the subject that is motivated by folk-psychological reasoning or the scientific strategy of compartmentalization of research problems. This project is both conceptual (subject and consciousness are rather theoretical concepts, hence their relation to more readily observable phenomena needs to be clarified) and empirical (because we want to understand the place of consciousness in nature).

1.3 Overview of the following work

The dissertation proceeds as follows.

In chapter 2 I describe the various meanings in which consciousness is said to be unified, and specify the meaning of the unity the explanation of which is pursued later in this work. The selected meaning is that of the unity of conscious contents at a time, i.e. their integration in a single coherent perspective of the world. I also propose that the so-called subject unity of consciousness, i.e. the

[13] I don't mean to suggest that this picture is inevitable, let alone the only conceivable one. But I do mean to say that where this picture prevails, understanding the subject's place in nature is consequently considered to be a greater problem.

explication according to which the unity of consciousness means that the contents belong to a single subject which unifies them, is inextricably tied to the first meaning and hence should be pursued in parallel. The reason for focusing on these two senses of the unity of consciousness is that they represent the greatest challenge for the scientific explanation of consciousness, and that conceptual confusion associated with them often leads to versions of Cartesian materialism.

Chapter 3 tries to clarify the main conceptual issues faced by naturalistic accounts of the unity of consciousness. I argue that the concept of the unity bears a normative and an objective aspect. The normative aspect consists in the fact that in attributing conscious contents to a third person we assume that the contents must be coherent, and that the agent's actions follow instrumental rationality. The importance of this consideration is that any empirical research of consciousness involves inferences about subject's conscious states that follow these constraints. The objective aspect is the proper domain of cognitive neuroscience: how the vehicles of conscious contents are integrated to underlie unified consciousness. I argue that an objective account of the unity of consciousness, that is explanation of the integration at the neural level, is needed because it is not possible to account for the unity solely in terms of conscious contents (as in accounts which argue that the unity is only represented).

Another conceptual issue dealt with in chapter 3 is the notion of neural representation. Virtually all scientific theories of consciousness employ the concept of neural representation. In contrast, the concept is used only reluctantly in contemporary philosophy after the criticism of computationalism and representational theory of mind. A large part of the chapter is thus dedicated to clarifying in what sense neuroscientists speak about representation and to showing that the concept is still useful for the discussion of the unity of consciousness if used with caution. A key part of this is understanding that the technical notion of neural representation does not directly map onto personal-level contents.

In chapter 4, I turn to interpretations of Kant's transcendental psychology in order to analyze the unity of consciousness further. In his *Critique of Pure Reason*, Kant provided a very detailed and insightful account of the unity under the term transcendental unity of apperception. The point of the chapter is to gather

ideas that link cognition, situatedness, consciousness and the unity thereof. I do not intend to interpret Kant myself, nor argue which interpretation is correct. The key findings of this chapter are the distinction between empirical and transcendental self-consciousness, the idea that perspective and hence subjectivity comes with the distinction of how things are and how they seem to be, and the idea that the unity can be understood as a result of the synthetic activity of the mind which in turn can be described in terms familiar to cognitive science. Kant's work also helps to articulate some misconceptions of the unity and shows that subject unity and integration unity as defined in chapter 2 are two sides of the same coin.

Chapter 5 presents a set of philosophical accounts of particular features of the unity of consciousness. First, I review Shoemaker's and Castaneda's accounts of the logic of the first personal pronoun 'I'. They show that the 'I' as a definite description cannot be reduced to a context-free description and consequently that we need a psychological account of self-reference without identification. Next, I summarize Hurley's two-level interdependence model of the unity of consciousness with special emphasis being put on her account of perspectival self-consciousness as constituted by egocentric action-perception feedback loops. Finally, I present an account of empirical self-consciousness where I put together the threads explored in previous chapters. The main point of this last preparatory part is to argue that the self-reflective state should be understood as representational transformation of the object state, not as a higher-order state that contains the object state as its proper part.

In chapter 6 I finally review three influential neuroscientific theories of consciousness: the neural global workspace theory, the predictive coding theory, and the information integration theory. I specify what the unity of consciousness amounts to according to these theories and assess how good their explanation is. The assessment results in favor of the predictive coding theory - not because it provides a decisive explanation, but because the predictive processing framework is congenial with many Kantian themes related to the unity of consciousness, it provides a compelling account of self-reference without identification, and it supports embodied cognition rather than the controversial representationalism.

Chapter 7 concludes.

2 Clarification of the concept of the unity of consciousness

Many meanings may be attributed to the concept of the unity of consciousness. This section will describe them without arguing whether or how consciousness is unified in that respect.

There are three dimensions along which one can make distinctions regarding the unity of consciousness: 1) temporal - whether the unity is considered as the unity at a time or over time, 2) qualitative - whether the unity concerns access or phenomenal consciousness, and 3) structural - what are the elements that form the unity.

2.1 Synchronic and diachronic unity of consciousness

Along the temporal dimension, a distinction is made between the unity of consciousness at a time (synchronic unity of the current state of consciousness) and over time (diachronic unity of successive states of consciousness). Synchronic unity of consciousness concerns the unity of a multitude of contents (or qualities, if one wants to avoid talking about contents and prefers phenomenal properties as the right level of analysis) of which we are conscious at a moment. Diachronic unity refers to that what makes a succession of conscious states part of a single, continuous stream of consciousness.

In terms of structure, synchronic unity is, broadly speaking, a matter of relations among currently represented contents, while diachronic unity is a matter of relations among the whole conscious states (the totality of the content of consciousness at any given moment).

Diachronic unity is often thought to be constituted by long-term memory of personal history. In this sense, the unity of consciousness is secured by a subject having access to the memory of her previous conscious experiences. We could say then that diachronic consciousness is a matter of representing and ascribing succession of conscious states to a single empirical subject in which they are thus unified. In this sense, diachronic unity is represented. Note, however, that the stream of consciousness, as we experience it, is not just a memory of isolated states. It includes a sense of their transition from one to the next. It is this sense that makes us think of consciousness over time as a continuous stream.

The need to account for this continuity thus invites a third possible category, namely the unity of consciousness over a short period of time. This kind of temporal unity of consciousness was famously analyzed by Husserl (2013) where he introduced the concepts of retention and protention to account for our ability to retain (and foresee) moments that just passed (and are just about to happen) in one specious present, and thereby constitute the sense of continuous transition rather than discrete succession. This kind of temporal unity is different from the previously described diachronic unity in that it is not represented. The unity of retention, protention and the present moment is not constituted by representing their corresponding contents as successive states of consciousness of the same empirical subject at different times. This temporal unity requires some special kind of co-consciousness relation (or synthesis, in Kantian terms). I shall refer to this kind of temporal unity of consciousness as synchronic*.

While it is likely that, empirically, any state of consciousness at a time involves synchronic* unity and there might not be a state having only synchronic unity, these two are conceptually different and we should hold them separate. The difference lies in the kind of relations (or syntheses) that are putatively at play. The constitutive relations of synchronic unity are atemporal.

2.2 The unity of access and phenomenal consciousness

In discussions of the unity of consciousness, it is sometimes specified that whether the unity concerns relations among phenomenal properties of conscious states or their represented content. There are many ways in which this distinction

has been articulated: subjective vs. objective character of mental states, qualia vs. functional properties (Chalmers), what-it-is-likeness vs. intentional states (Nagel), or phenomenal and access consciousness (Block).[1] Hereafter, I will use the distinction between phenomenal and access consciousness.

In his seminal paper, Block (1995) introduced the distinction using Nagel's phrase of what it is like to be in a state to characterize phenomenal consciousness. In contrast, access consciousness is characterized in terms of a mental representation directly influencing one's behavior in virtue of the information it carries. More specifically, access consciousness is delineated by a set of three sufficient (but not necessary) conditions:

> A state is access conscious (A-conscious) if, in virtue of one's having the state, a representation of its content is (1) (...) poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech. (Block, 1995, p. 231)

Definition of phenomenal consciousness is more elusive, as Block himself admits. It is supposed to refer to experiential properties of our mental states that are independent of their intentional and cognitive properties. It is not clear whether this independence is held to be empirical or just conceptual. However, the interpretation of the motivating case for Block's distinction, namely blindsight (as a a case of access consciousness without phenomenal consciousness), suggests that Block thought that access and phenomenal consciousness can be empirically independent of each other. On the other hand, Chalmers (1997) argues that the two are perfect correlates and the distinction is therefore conceptual.

2.2.1 Access consciousness: integrated representation

According to Block's definition, the unity of access consciousness would correspond to the integration of contents in joint control of behavior: what must the organization of conscious representations be like so that they all are poised to jointly control action or reasoning? Given that access consciousness refers to

[1] Although the distinctions are different, the sense of what they try designate at the subjective, qualitative level as well as the objective, content level is similar enough to consider them as a single distinction for the purpose of the present discussion.

the representational and functional properties of conscious mental states, its unity can be conceived of as a matter of integration of representations or functions such that the organization constitutes a recognizable whole - a person, for example.[2] This characterization goes further than that of Bayne and Chalmers (2003) who describe access consciousness as:

> two conscious states are access-unified when they are jointly accessible: that is, when *the subject* has access to the contents of both states at once. (Bayne and Chalmers, 2003, p. 8, emphasis added)

The more complicated characterization should be preferred because it avoids reference to an independent subject - for if we refer to a subject in our description of what the unity amounts to, the question about what constitutes the unity may seem trivial (for one could argue that talking about a single subject already assumes the unity and hence that there is no problem of the unity itself - above and beyond the problem of what constitutes a rational agent). The unity issue is non-trivial if we ask how mental representations and functions are integrated to constitute a single rational[3] agent. This, in essence, is what researchers investigating high-level mental functions like reasoning or decision-making study.

To be fair, many philosophers probably think that empirical research of how such functions are realized can only reveal what (physical) conditions must be met for a system to behave in such a way that we then understand it as a single agent - but it won't tell us anything about how the experiencing subject (from the first-person perspective) is constituted. Naturally, the experiencing subject is then considered to be an inexplicable assumption of all studies concerning

[2]At the most basic level, the evidence for integrated representation is such a system's (agent's) action that it is arguably driven by heeding to separable representations. For example, an agent that explores two options and then chooses the better one can be said to have an integrated representation of the choice (comparison of the two options). Of course, to the extent to which the action can be interpreted as driven by simpler, non-integreted representations, or even urges and associations, the case for integrated representation and hence access consciousness is weaker. But this is a drawback only if we take seriously the possibility of philosophical zombies that can behave rationally without being conscious.

[3]Rationality and the normative aspect of the concept of the unity of consciousness is discussed in greater detail in 3.

consciousness; and referring to the subject in definitions of consciousness-related phenomena is consequently seen as endorsing this assumption rather than falling to an infinite regress. Since the goal of this work is to assess scientific accounts of the unity of consciousness, I must assume that third-personal accounts of what constitutes the unity may be relevant for what constitutes the subject *per se.*

Studying the unity of access consciousness, as opposed to phenomenal consciousness, has a clear advantage in identifying the relata of the unity relation, namely representations (or mental states, if one wants to avoid any kind of commitment to a representational theory of mind). This is not to say that there is no ambiguity in telling how much detail a particular mental state represents or how it is in fact represented. But if we do talk about mental states (use them as theoretical entities in our explanations), we identify them in virtue of their intentional object or the function they serve. Note that this identification is indispensable even when discussing P-consciousness of mental states - for even if it is conceptually possible that phenomenal properties of a mental state are not type-identical with its functional or representational properties, the mental state itself is still individuated by the latter.

2.2.2 Phenomenal consciousness: one experience or many?

Although I don't think that the concept of phenomenal consciousness is fruitful for the unity problem, I will, for the sake of completeness, discuss what the unity of phenomenal consciousness amounts to.

If access unity is a matter of integration of represention (the domain of access consciousness), a charitable interpretation would be that phenomenal unity refers to the way mental contents are integrated in the experiential dimension. That is: how phenomenal property of a complex conscious state relates to the phenomenal properties of the constituents of that state. According to Brook and Raymont (2017), there are two alternative views: either one believes that there is only one, total experience at a time, or that there are many experiences constituting the total experience. The experiential parts view conceives of the unity of phenomenal consciousness as a matter of relation among phenomenal properties of the component mental states. Thus the experiential parts view aims to explain

how a multitude of experiences (with their phenomenal properties) is unified into what it is like to experience their combination. One such view is proposed by Bayne and Chalmers (2003) and Bayne (2010) where the relation at stake is called subsumption. The non-experiential parts view holds that there is only one experience at a time, i.e. one global state of which we are phenomenally conscious. Although there often are many objects of which we are conscious, there is only one state with the what-it-is-likeness property that *is* the total P-conscious state.

The trouble with both alternatives is that neither makes a good sense, at least if we want to understand what constitutes the unity, and not just describe what it means. On the non-experiential parts view, the unity follows trivially (there is only one state). The challenge for this view is to explain the relations of what-it-is-likeness of total conscious states that share most of the content or the change in what-in-what-is-likeness when new content enters consciousness. To say that a new P-conscious state arises everytime the content changes is utterly non-informative regarding the unity of consciousness because the latter is then stipulated rather than explained. The experiantial parts view faces a reversed problem: unless there is a way to identify phenomenal properties of mental states without referring to their content (and I don't know of any such way), there is a problem of how to conceive the phenomenal property of a complex mental state (composed of, say, states A and B) other than as the phenomenal property of the union of A and B. That is, talking about phenomenal properties on top of representational ones would be justified if the phenomenal properties did not have one-to-one correspondence with representational properties. But how could it not if phenomenal properties are identified in virtue of the states' content? Of course, some may be inclined to accept some version of parallelism of representational and phenomenal properties, but in so far as the unity is concerned, focusing just on the representational side will be enough.

2.3 The unity of conscious states

The last dimension (structural, as I call it) invites many distinctions, depending on how fine-grained a classification of mental states one prefers. We can think of distinctions in this dimension as answers to the question "Into what are

conscious representations unified?" For our purposes, we can, with some modifications, follow the distinctions put forward by Bayne and Chalmers (2003). They distingiush 1) the objectual unity, 2) spatial unity, 3) subject unity and 4) subsumptive unity.

1. The objectual unity is a matter of ascribing properties to a common object. In cognitive science, this is known as binding - integration of feature detectors in lower levels of the information processing hierarchy. The traditional (representational) view is that the object is then recognized as something (e.g. as a phone or a cow) thanks to the recognized features. It is not necessary, however, that the object falls into a definite semantic category - it is plausible to assume that we can bind features of an unrecognized object, i.e. represent them as coming from the same source. In other words, objectual unity is strictly speaking a matter of identification (attributing features as belonging to an identical object) rather than recognition, although the latter often follows from the former.

2. The spatial unity is a matter of locating objects (including sounds etc.) in space. From the representational perspective, space is not a special object. Rather, it is a framework in which objects can bear spatial relations to one another. Or, as Kant puts it, it is an a priori form of intuition. Describing the spatial unity like this does not make any strong ontological commitment to the existence of a special framework on top of spatial relations themselves: one can think of space as constituted by the relations. Following this logic, one could perhaps argue that temporal unity could be understood in this way as well, but spatial and temporal unity have usually been held separate.

3. Subject unity simply states that various representations are unified in a single subject - the subject that is having them. Note that according to this formulation, the representations do not need to entail, as part of their representational content, that they belong to the same subject (as would be the case in acts of self-reflection, for example). Insofar as the subject is understood just as an abstract placeholder to which mental states are attributed from a third-person perspective, subject unity seems to be a

trivial statement following from the fact that consciousness is, by definition, always someone's consciousness. In this sense, the subject has a similar theoretical function as, for example, the center of gravity. To illustrate this, we can speak of internal cognitive states of robots or programs and thereby constitute a subject - as a placeholder for a system that has a limited amount of information at its disposal, or acts in a way that warrants adoption of intentional stance towards it. What makes subject unity non-trivial is that consciousness, at least in our case, comes with self-awareness. That is, the subject is something that we know from the first-person perspective to be more than just an abstract placeholder. Clarification of this point is notoriously difficult and its proper discussion is postponed to later sections. The act of self-reflection provides a sense of identity of the reflecting and reflected-upon subject.

4. Subsumptive unity is a name that Bayne and Chalmers give to probably the most common meaning of the statement that consciousness is unified. The meaning is the unity of all current representations in a single state of consciousness. This single state of consciousness is sometimes metaphorically referred to as conscious field. The point of this concept is to bring attention to the intuition that our conscious experience at a time is not just a set of unrelated conscious representations but rather a whole encompassing mental states which are so different that they do not share any obvious integrating framework, such as space is for visual perception. At this moment, I am conscious of the proprioceptive feedback from my typing fingers, of various alternative expressions that I could use in this sentence, of a persistent neck pain, of a lunch aftertaste in my mouth and noises generated by people working in the office. We say that these various representations are unified in one consciousness (as opposed to some lower-order framework, such as space). Bayne and Chalmers call this unity subsumptive because they explicate it in terms of a specific relation among conscious contents: "two conscious states are subsumptively unified when they are both subsumed by a single state of consciousness."[4] The subsumption architecture is

[4]Bayne and Chalmers (2003)

not the only way to specify the relation among component representations, however. So, I will use the term 'integration unity' to refer to the general fact that at each moment our conscious experience comprises contents or representations from different domains or senses that are related to each other in some way.

2.4 The unity analyzed

Having outlined the various meanings attached to the concept of the unity of consciousness, I can now specify which one I will pursue. In the temporal dimension, I will focus only on synchronic* unity because the diachronic unity of consciousness is a very different issue related to personhood rather than consciousness as it is understood in the philosohpy of mind and cognitive science. In the qualitative dimension, I will focus on access consciousness because my aim is to explicate the unity in terms of structure of representations or information processing; and phenomenal aspects of mental states are held to be conceptually independent of their content. Finally, in the structural dimension I will focus on the subject unity and the integration unity as these two 1) pose the greatest challenge in naturalizing consciousness, and 2) are so strongly intertwined that one cannot be explained without the other. In arguing for these two points I turn to Kant's analysis of the necessary unity of consciousness.

3 Normative and Objective aspect of the unity of consciousness

Having described the meaning of the concept of the unity of consciousness, I would now like to discuss and clarify its relation to other concepts as well as address some methodological issues that arise when we set out to study the unity empirically.

Let's begin by stating that the unity is an essential property of consciousness. This claim invites two interpretations.

The first is that we cannot make sense of a consciousness that would not be unified. When we ascribe consciousness to a person (or animal, or a system), we thereby ascribe to her the capacity to integrate information at her disposal and, on that ground, to act rationally. This is the normative aspect of the unity of consciousness: rationality is a norm that we apply, often implicitly, to assess the extent of someone's consciousness (the extent of what is unified).[1] For example, if we have good reasons to believe that a person is thirsty and we observe that she does not take a sip from a glass of water that is clearly in her line of sight, we conclude that the person is not conscious of the glass - the more so if we have a good reason to believe that something is wrong with her visual system, as in the case of hemi-neglect.[2] The rationality assumption may not always adjudicate univocally the question of what conscious content to ascribe to someone. For example, in the case of split-brain patients, the capacity for rational action based on integrated information is the criterion that leads Nagel (1971) to speculate

[1] In the same vein, Dennett (1989) holds that agent's rationality is a key assumption in adopting intentional stance towards it.

[2] Note that we could as well conclude that she is not conscious of being thirsty, though intuitively the idea that one could be thirsty without realizing it is more odd than the idea that one could fail to see something while looking at it.

about two centers of consciousness, each pertaining to one hemisphere. Regarding the same set of cases, Gazzaniga (2000, 1985) is led by this norm to the conclusion that perhaps only the left hemisphere, responsible for language production, is conscious, thus speculating about the prominent role the left-brain interpreter may play in healthy brains in explaining consciousness..

The norm of unified consciousness manifested in rational action is also important in discussions of animal cognition. For example, experiments of Cheng (1986) show that rats are unable to integrate olfactory and spatial representation to solve a maze. In his experiment, food was hidden in one of the corners of a box. The location could be deduced from olfactory markers and the shape of the box, each reducing the number of alternatives to two. Although the combined geometrical and olfactory information reduced the number of alternatives to one, the rats were still searching for food at the two locations indicated by the shape of the box only. Given that rats have a strong sense of smell, their inability to use the olfactory information for rational control of action is interpreted as a lack of integration, thus narrowing down the extent to which rats can be said to be conscious.[3]

The second interpretation of the unity as an essential property of consciousness is that consciousness could not serve its evolutionary function (roughly: enable the organism more flexible, context-sensitive reaction to changing environment) if it did not unify various contextually relevant representations. This is the objective aspect of the unity: the unity is a constitutive feature of consciousness. Some empirical theories go as far as saying that consciousness *is* the integrated

[3]This is only to illustrate how the norm of rationality is applied to determine the extent of consciousness. Alternative explanation of the same experimental finding would be that searching at two possible locations was fast and easy enough not to drive learning to integrate the olfactory and geomatrical representations.

One could further argue that failing to narrow down the possible locations to a single one is perhaps a matter of attention or inferential capacity, rather than consciousness as integration. The second option can be rejected on the ground that solving the task requires simple association (as opposed to an inference allowing solution to a novel problem based on past experience) between food, geometrical shape and an oflactory marker. The first option can be rejected on the assumption that if a subject fails to utilize some information on repeated trials, it is because she is not conscious of it, not because she did not attend to it.

information.[4] It is this aspect what makes the unity of consciousness an empirical problem, not just a conceptual one.

Theorists sympathetic to the Hard problem formulation by Chalmers (1995) would probably disagree. Specifically, they would argue that philosophical zombies (or robots, to choose a more realistic scenario) with the kind of integrated information allowing for flexible action are still conceivable. Bayne and Chalmers (2003); Bayne (2010) hold that phenomenal consciousness is unified and they grant that information integration may be a necessary but not sufficient condition of phenomenal unity. It is not sufficient because it alone does not guarantee that there will be anything like to be this access-unified system - the metaphysical possibility of zombies purportedly demonstrates this. Thus theorists who have a strong zombic hunch, as Dennett calls it, will reject the idea that integration might be sufficient for consciousness.

Now, even if we don't accept the zombie argument, we may be reluctant to agree that the objective aspect of unity poses an empirical problem. Why? Employing the content/vehicle distinction, the argument would run as follows:

Talking about the access unity (unity in its objective aspect) implies discussing the unity at the content level, leaving aside how the accessible contents are realized or what they supervene on. The access unity as a necessary condition of consciousness requires that the agent can integrate various representations and manifests this by an inference (practical inference suffices). Since representations (contents) are mutliply realizable, and inseparable from the world,[5] no description of brain processes *as such* (as vehicles of representations) will tell us what contents are being represented and hence what is the principle of the unity at the vehicle level. Or, to put it broadly, consciousness manifests itself in the agent's interaction with the world, not in what the brain does. The idea that the unity is an empirical problem arises from conflating the content/vehicle distinction and assuming that because *vehicles* of conscious contents are brain processes, studying their causal relations will provide an explanation of relations at the content level.

[4]See chapter 6 for precise formulations of the identity.

[5]See Hurley (1998) for a thorough argument for externalism about content.

The rest of this chapter will try to defend against this argument and lend some support to the idea that the unity is not only a defining feature of consciousness but also an empirical problem. There are two lines of defense: 1) I will argue that neuroscientific explanations are not limited to causal accounts of brain processes *as such*; they may also describe the general processes by which brain states (vehicles of representations) come to function as representations (contents). 2) I will closely examine what warrants talking about representation (content) in the brain so that it is clearer how studying relations between representations at the vehicle level can shed some light on their relation at the content level.

Before I go into the details, let me add two points. First, as Hurley (1998) and Dennett (1991) show, it is wrong to assume that vehicles of a conscious content must have the same features as the content they realize. Obviously, a percept of a green object is not itself green, or less obviously, a represented synchrony of two percepts is not necessarily a matter of synchronic representings. Hence we must be careful not to take any principle of unity at the level of vehicles of conscious contents as constitutive of the unity of consciousness. Presumably, the vehicles of conscious contents are neural activations and some of the neuroscientific theories of consciousness are based on a single feature that the neural activations corresponding to conscious representations have in common (for example, gamma band synchronization, see Fries (2009, 2005)). The theories usually hold the feature in question to be functionally important, i.e. to underlie the integration of the representations into the whole of the current state of consciousness. The crucial qualification is, of course, "functionally important". Mere coincidence of a feature of neural processing with conscious states does not guarantee that the proposed common feature is the key to the causal explanation of the unity of consciousness.[6] The proposed feature that unifies some brain states must be relevant to the function of consciousness.[7] This is a reason why theorists denying that consciousness has a functional role are skeptical to the prospect of neuroscientific

[6]To anticipate, Tononi's integrated information theory is particularly liable to this objection.

[7]The justification of the claim that some feature of neural processing unifies neural representations would depend on showing that the feature realizes the functions of unified consciousness, not vice versa. It is not the case that any principle of unity at the neural level is *eo ipso* the constitutive feature of the unity of consciousness. As (Hurley, 1998, p. 39) puts it: "Of course,

explanation of consciousness. Without a recognized function of consciousness, all that neuroscience can offer are models based on neural correlates which are prone to be attacked by "zombic" arguments. Thus I will first summarize the functions attributed to consciousness in cognitive science. After that, I will try to clarify the concept of neural representation as it used in cognitive neuroscience. Finally, I will address the question whether the unity of consciousness is a matter to be explained by neuroscience.

3.1 Function(s) of consciousness

The project of devising a naturalistic account of consciousness can be convincing only under the assumption that consciousness, like other biological adaptations, has some adaptive function(s).[8] Only then can we hope to demonstrate that a particular set of neural processes constitute consciousness by fulfilling the adaptive functions in question. Although we can find neural *correlates* of conscious experience even without such functional description of consciousness, these correlates would always be open to the question "Why do these correlates give rise to conscious experience?" Naturally, those who have a strong zombic hunch will argue that what is interesting about consciousness cannot be captured by any functional description, thus securing applicability of the open question argument to any naturalistic account proposed. For the rest, however, the functional description will serve as that against which the adequacy of a proposed account can be measured. So let me elaborate a bit on the functions of consciousness.

The canonical functional description of consciousness comes from Baars (1988) who lists 9 major functions. In summary, they cover four broader areas: executive control (decision making, agency), learning, self-monitoring, and sensitivity to context.

The executive role of consciousness is prominent in cases of conflicting goals that need to be prioritized to resolve the conflict and act consistently. Regarding

when we find out what the functional basis for unity is, we can call that a kind of unity relation among vehicles. But it may or may not have any independently identifiable unity."

[8]Note that holding this view does not entail commitment to classic functionalism which is a theory of what constitutes contents of mental states.

specifically the unity of consciousness, the aspect of executive control worth noting is that of bringing various (potentially conflicting) goals together and assembling them in a hierarchy conducive to consistent behavior.

As for learning, there is a lot of evidence that learning a novel task requires conscious pondering of the task-related features. The reason why consciousness facilitates learning a novel task is that it allows formation of a new context that needs to be actively maintained. For example, learning to drive a car requires one to attend to the visual feed, manipulate the car with hands and feet, tapping onto one's knowledge of traffic rules, and be aware of the traffic situation. Until we learn the common patterns of translating information from multiple domains to an appropriate reaction (e.g. what to do when we approach an intersection without a STOP sign and there is a car coming from the left), we need to maintain the relevant information in consciousness to take the right action. (From information perspective, consciousness helps us navigate through the initially large state space until we learn the usual patterns by repeated practice and thereby reduce the state space to its more manageable subset. From then on, automaticity usually takes over.)

Self-monitoring and metacognition are functions that I treat at greater detail elsewhere (3.2.3, A).

Finally, context sensitivity is an area that has already been echoed in the first two. According to Baars, this is the main function of consciousness: combining various sources of knowledge to form a coherent experience that enables the conscious agent to act adaptively even in novel situations that she could not have rehearsed before. The contrast is with reacting habitually to a salient feature in the environment, not taking into account the context of the situation - for example, running away from a lion that is safely locked in a cage (despite being motivated to stay). Context sensitivity thus involves integrating information from the environment with previous experience and the hierarchy of one's goals.

3.2 Representation in the brain

There seem to be good reasons to be skeptical about claims that the brain realizes cognition in virtue of representing things in the world and their proper-

ties. To see why, consider some arguments from early discussions of connectionist models of cognition. On the assumption that the brain realizes cognition in much the same way as connectionist networks, Ramsey et al. (1990) argue for eliminativism of folk-psychological concepts, specifically propositional attitudes. Their argument starts by analyzing folk-psychological assumptions about mental states: they are supposed to be 1) functionally discrete (beliefs are held to be functionally independent), 2) semantically interpretable (about something), and 3) causally efficacious (they cause behavior). They then argue that nothing in connectionist networks satisfies these assumptions. The main reason is that beliefs or representations that purportedly underlie the execution of the task the network is trained for may be attributed only to the whole network, not to a part of it. As the network learns and changes its output in a way that can be described as a change in beliefs or representations in the hidden layers, what really changes is the whole set of weights, not a distinct subset.[9] In other words, representation and computation is distributed over the whole network. Naturally, if we cannot identify individual representations *within* the network, we cannot say that some representations *individually* cause some action (in the way that folk-psychology holds that the belief that it will rain causes, together with other beliefs and states, taking an umbrella), nor can we say that some part of the network is about this rather than that.

Against this view, Smolensky (1995) argues (rightly, I think) that Ramsey et al. (1990) omit the possibility of identifying the attributed representations with higher-order features of the network, and provides few examples of post-training analysis of higher-level network features that may be more plausibly identified with representation. Moreover, the arguments against connectionist

[9]Actually, even if only a part of the network changed, we could not say that the changed part corresponds to the changed representation, for the output that hypothetically follows from that belief would still be functionally dependent on the unchanged weights. Only if we could manipulate the unchanged weights without affecting the output that hypothetically follows only from the attributed representation, could we say that the representation is realized by the weights changed due to learning. And even in that case it would be highly likely that the changed weights are implicated in other attributed representations. Thus the condition of functional independence is not satisfied by connectionist networks.

representations are based on fairly simple networks - simple both in function and scale. Functionally independent representations are more likely to be found in more complex networks, such as the brain, that may perform a great variety of tasks that consequently rely on unrelated, and hence modularized, representations. In the rest of this section I will show that in looking for representations in the brain, cognitive neuroscience relies on the same idea as Smolensky advocated, namely that representations can be identified with higher-level features of neural networks.

3.2.1 Neural correlates

First, it will be useful to describe few examples of neuroscientific explanations to see what claims about neural representation actually amount to. Classic examples often come from visual processing as vision is the most thoroughly studied sensory modality in neuroscience. Neurons or neural circuits[10] in visual cortex act as feature detectors for such simple things as oriented edges at a specific part of the visual field. That a particular neuron is representing, for example, a line tilted at 50 degrees is inferred from its selective activity (action potentials, or spikes) to such lines. The relation between stimulus and neural response can thus be described by a so-called tuning function that maps stimulus feature space onto neuron's firing rate. The neuron would spike most frequently in presence of lines tilted at 50 degrees, less frequently for lines at slightly above and below that angle, and would be silent in other cases. Actual orientation detecting neurons always belong to a specific receptive field (part of the visual field) and can possibly code for other things as well.

Importantly, what a neuron is selective to is always determined only to the extent experimental manipulation varied conditions. If, for example, it turned out that a neuron was spiking at the presence of a 50-degree *blue* line and did not spike at stimuli of other colors and orientations, we would infer that the

[10]To keep things simple, I will assume that single neurons can code a feature. In fact, even simple features are usually coded by a group of neurons. The simpler the feature, the more localized the neural group coding it. The following argument holds for neural circuits, the difference being only that the activation pattern which correlates with the presence of a stimulus is more complex.

neuron codes for co-occurrence of a color and an orientation. In this sense, the proposed neural representation is always underdetermined as further experimental manipulation might show greater selectivity (or wider sensitivity) of the neuron or neural population.[11] As Dehaene and Naccache (2001) note, correlational evidence for neural representation is more robust if the correlation holds even in non-trivial cases such as illusions or mental imagery. For example, perception of motion caused by some visual illusions is correlated with neural activity in V5 - the same area that is active during perception of non-illusory motion of physical objects.

It could be argued that the concept of neural representation is further compromised by the fact that it almost never is the case that a single neuron codes a single feature. The idea, known in neuroscience as the grandmother cell hypothesis, that there is a particular neuron that is uniquely active in the presence of things falling under a specific concept (e.g. grandmother), is widely agreed to be false. Such an arrangement would not be very adaptive given that thousands of neurons die every day. It is thus thought that the brain employs sparse coding - a feature is coded for by a group of neurons that is relatively small compared to the number of all neurons in the brain but large enough to avoid loss of representational capacity due to natural decay. Sparse coding thus allows for a concept

[11]There is a notable similarity to Quine's argument on the indeterminacy of translation. As 'gavagai' could be translated, given evidence, as both 'a rabbit' and 'an undetached rabbit-part', so could any neural activation afford many interpretations of its represented content given the necessarily limited experimental evidence. An illustrating example of this problem in neuroscience is the interpretation of the function of the fusiform face area (FFA). As its name suggest, it had been long thought that this area is specifically involved in face recognition because it was distinctively activated in the presence of facial stimuli. Researchers reaching this conclusion used pictures of ordinary objects as the control condition. Later, I. Gauthier and others showed that the FFA responds not only to faces but to any *specific* object (not generic, i.e. not a chair, but the old chair I always sit on while working) because they used a different control condition. The FFA thus serves to distinguish individuals within the same category where it is needed - which is often the field of expertise of the particular person (e.g. various chess positions for professional chess players). Since the area of expertise varies greatly among subjects, this relationship had been difficult to spot experimentally. The reason why the FFA is active in any person's face perception is that distinguishing indvidual people is an area of fine discrimination that we all have good ecological reasons to be experts in.

to be coded in such a way that malfunction of few neurons will not result in representation of something else completely but rather in a representation of a semantically similar concept, e.g. LAKE and POND. In this view, each neuron ought to be part of multiple groups coding for specific features - otherwise the range of possible representations would be very small. It might thus seem that sparse coding and distributed, overlapping patterns of neural activations undermine the possibility of meaningful identification of represented content and brain processes.

On closer look, however, it only undermines the idea that representation can be neatly *localized* in the brain. The example above shows that the brain has at least one way of representing oriented lines. Furthermore, where technology allows, neuroscientist always try to go beyond mere correlation, i.e. they try to manipulate neural activity directly and see whether it elicits a response that is expected on the hypothesis about what the neuron (neural population) is coding for.

Generally, there are multiple levels of evidence that the brain represents some information. The first, most superficial level would be showing how certain task that the brain is capable of performing can be accounted for at the computational level in terms of intermediate representations and their transformations. A generic example of such stage are information processing diagrams and models typical of classical cognitive science, e.g. Baddeley's working memory model. The models are designed so as to accomodate the bevioral evidence, typically all the systematic mistakes subjects make and the unusual impairments found in neurological cases. The more of behavioral evidence is accumulated (often with the direct aim to test and disprove a model), the more complicated the models become and the more representational stages are hypothesized. Next level of evidence would be finding neural correlates of the hypothesized intermediate representations. This can be done to various degrees of precision. An example of a crude correlate would be a distinct EEG signal occurring only in cases where the investigated representation is hypothesized at the computational level (so-called event-related potential). An example of a fine correlate would be a result of voxel-wise modelling in fMRI where the BOLD signal is fitted to a non-linear

transformation of hypothesized feature space. For example, Nishimoto et al. (2011) investigate what features of an image the brain uses to represent visual scene by modelling the brain's activity during watching video clips, using various algorithmic transformations of the presented images as independent variables. One such transformation, for instance, contains only information about areas of high contrast in the image. If an area is found in the brain whose activity correlates with this numeric measure of contrast, it is inferred that it codes and hence represents this visual feature.

Further level of evidence is provided by manipulating neural activity and observing corresponding changes in behavior or reported conscious representation, or from knowing how one representation-type is transformed into another. For example, we now know that the location of a sound source is partly coded by the interaural time difference (sound waves arrive in the ears at slightly different time, unless the source is directly in front of us or behind us). The difference is represented by neurons that function as coincidence detectors: they receive signals from primary auditory cortices for both ears and spike only when the signal comes from both ears at the same time. Unequal lengths of the presynaptic axonal projections ensure that signal from one ear travels to the difference-coding neuron faster and thus compensates for the difference in time at which the sound arrives at the ears.[12] If we could delay the signal coming to the coincidence detectors and if we observed a change in the estimated location corresponding to the delay, we would have strong reasons to believe that the brain indeed represents location of a sound source by means of interaural time difference at the algorithmic level, and by coincidence detectors with presynaptic connections of unequal lengths at the implementation level.

3.2.2 Neural representations

A skeptic could still argue that such causal demonstrations only show that the manipulated neurons are implicated in the response, not that they represent information on which the response is purportedly based. That objection is misguided, I think, because for cognitive neuroscientists mental representations are

[12] See Carr and Konishi (1990) for further details.

theoretical posits that have a useful explanatory role, not something that has the representational status *per se.* In other words, that neural activity represents information is a central theoretical assumption of cognitive neuroscience: complex behavior is considered to be a result of information processing between inputs and outputs. In this paradigm, the process is ideally first described at the computational level (what is the task the organism faces and what kind of information it could use to solve it). In the next step, scientists look for processes corresponding to the hypothesized information. If the same processing structure is found at the neural level as was hypothesized at the computational level, it is often stated that the brain represents that information. This, however, is too loose a formulation. A more precise statement in this case would be that the brain performs a task by a mechanism interpretable along the description at the computational level.

What motivates the more stringent formulation is the idea that not every correlate between a stimulus and a neural activation should be called a neural representation. There are many cases in which a neural activation is *just* a causal mediator of a physical response - for example an activation of a reflex arc. To say that the activation responsible for the gag reflex *represents* the danger of a solid object entering the throat would be too loose, for then any (neural) effect would represent its cause. The concept of neural representation thus must be specified beyond mere correlations. There are two conditions that, in my view, render the concept of neural representation applicable but not too permissive.

The first condition is modular, detachable use: a neural correlate is a representation only if it leads to different behavioral effects depending on which module is making use of the correlate. This rules out the cases in which the activation is just a causal mediator of an action. For example, a pattern of activation in the motor cortex that correlates with a physical movement can be interpreted as a representation of that movement because the same pattern occurs if the agent is executing the movement as well as when she sees it or imagines it. On the other hand, a correlate of the movement's execution in cerebellum is not a representation because other modules do not use it (it does not occur in any other but the execution case). That a representation can be used by multiple modules can be

in principle ascertained by experimental manipulation - if we change the activation one way, the behavioral effect should change according to which module is making use of the representation.[13]

The second condition, or rather a constraint, is that a correlate can be meaningfully interpreted as a representation only if we show that the organism employs the correlate for rational control of action that cannot be explained more parsimoniously than as a manipulation of representation - for example, as a simple association. The interpretation of experimental results always entails the assumptions that the subject is rational and intends to pursue the goals that the researchers implicitly attribute to her (most notably to comply with experimental instructions in case of human subjects, or to get a reward in case of animal subjects). Without these implicit assumptions it would be impossible to differentiate between cases in which the agent acted in two different ways because she had different representations from cases in which she had different intentions. What a neural correlate represents is then relative to the context of intentions that we attribute to the subject, and to what we recognize as a rational action.

The underdeterminacy of neural representation with respect to experimental design is a corollary of the idea that content is constituted by patterns of interactions between an agent and its environment, not just by relations among internal cognitive states. (Haugeland, 1990, p. 386) argues that any account of content (intentionality) must be holistic in the sense that "the intentionality of any individual state or occurrence always depends on some larger pattern into which it fits",[14] for nothing can represent something else solely in virtue of its physical structure. He then identifies three main approaches to intentionality based on what kind of pattern they take to be constitutive of original (as opposed to derivative) intentionality. Neo-cartesianism holds that mental states have their contents in virtue of their systematic relations to one another, neo-behaviorism holds that the relevant pattern is that of agent-environment interactions, and finally neo-pragmatism appeals to patterns of normative practices realizable only within a community of social agents. It is not hard to see that *practically* speak-

[13]The commitment to a modular view of the mind is inevitable - the concept of neural representation is meaningless without it, I think.

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

ing the most appealing approach to intentionality for cognitive neuroscience is the neo-behaviorist view. Even if neo-cartesianism was right and the content of mental states was constituted solely by syntactic (systematic internal) properties of the brain states, mapping the patterns of interactions among brain states is technically much more difficult than mapping systematic relations between brain states and environmental stimuli. And on the other side of the spectrum, neo-pragmatist account of content leads to the conclusion that only personal-level mental states represent (have content) and hence that it does not make sense to interpret brain states as representational. In addition, neo-pragmatist emphasis on the social meta-norms of conformity and censoriousness renders studies of neural representation in animals (including not only the quite social primates but rats as well) mostly irrelevant for understanding how humans represent things. The alliance of neo-behaviorism and cognitive neuroscience is further cemented by the nature of information processing that is characteristic of neural processes: it is parallel and distributed (serial processing of formal symbols, on the other hand, is the domain of the neo-cartesian view).

So, I conclude that there are good reasons to talk about neural representation. Note, however, that the criteria mentioned above tell us nothing about whether the representation is conscious or not. Also, nothing of the above suggests that type-physicalist reduction of mental states is possible throughout. There may be types of mental states for which no reduction to types of brain states is possible. Notably, the simpler the represented feature, the more likely it is that we can find a localized neural correlate (e.g. oriented edges, tones, and perhaps faces). For more complex contents, however, it may be quite unlikely that we find their neural correlate (even intra-individually, let alone across individuals whose brains may differ considerably in the idiosyncratic connections underlying the same function).

Finally, it is worth dinstinguishing the question about neural representation from the issue whether folk-psychological concepts such as memory, consciousness, attention, etc. refer to anything even remotely homogeneous that could be found at the neural level. The latter is the crucial question of eliminativism and various authors, most famously the Churchlands, argue that cognitive neuroscience would make faster advance if it stopped trying to explain folk-psychological con-

cepts. More importantly for our case, Irvine (2012) argues that consciousness is not a good scientific concept. She reviews the usage of the concept in cognitive science literature and concludes that not only is there no convergence of empirical studies of the neural underpinnings of consciousness, but there is not even an agreed-on operationalization of consciousness. Similarly, Francken and Slors (2014) argue that the apparent lack of convergence of neuroimaging results for much of psychological concepts is a consequence of undue insistence of cognitive scientists to describe the results in terms of folk-psychological concepts which individual researchers often understand, and hence operationalize, differently. In contrast, looking for neural representation may not be informed by folk psychology only. As the research that tries to decode visual processing of video clips shows (see p. 35), the represented features were hypothesized on computational, not introspective grounds. Computationally defined representation has a much clearer definition and its operationalization is straightforward.

3.2.3 Neural metarepresentation?

If neural representation is problematic for the reasons sketched above, the notion of metarepresentation is even more so. It is not clear whether there are similarly reasonable criteria for neural metarepresentations.

Some theories of consciousness employ the concept of metarepresentation or higher-order representation that is, loosely speaking, about another representation, for example 'I think X' or 'X is veridical', where X is the object representation. First thing to note is that the metarepresentation is not a mere copy of the object representation, it transforms or adds some content to the object representation. Representing the same thing twice is metabolically costly and the only function it could serve is to have a backup in case one representational system fails. This, however, is probably solved more economically by sparse coding - a principle of representation allowing for graceful degradation of the represented content in case of partial impairment of the representational system (neurons, nodes in artificial neural network, etc.).

Now, if metarepresentation adds contents to or is a transformation of the object representation, is there a principled way to distinguish it from a represen-

tation that is just at a higher level of the processing hierarchy? For example, a representation of a particular face is based on lower-order representations of facial features such as eyes, lips, nose etc. At the neural level, insofar as we can distinguish the representation of the whole face from the component representations of eyes, lips, nose, etc., we can test empirically whether the neural activation corresponding to the latter gives rise to the neural activation corresponding to the former (plausibly with some feedback to the lower-order representation, as is most probably the case, see section 6.2). This seems to be a clear case of a higher-level representation synthesizing lower-level features. What makes us classify something as metarepresentation is that its content is *about* the lower-order representation *qua* representation, e.g. something is predicated of the representation, not of the represented object. But this is a distinction at the content level, not at the level of vehicles of content, i.e. patterns of neural activations. Hence we shall not assume that there is a difference between representation and metarepresentation at the neural level.

The reason why one might be inclined to think there is such a difference is that the formal description of a metarepresentational content, 'I think X', contains its object thought as a proper part. If metarepresentation consists of adding 'I think' to the object representation, could it not be the case that at the neural level it is realized by a specific physical relation between the object representation X and something that corresponds to the 'I think' that is constant across all metarepresentational states (say S)? No. First, despite the grammatical appearance, metarepresentation cannot be a result of mere syntactical operation of adding 'I think' to the articulated object thought. The reason why 'I think X' is an adequate formulation of the metarepresentational content is because the higher-order thought entails recognition of oneself as the subject of both the object thought and the higher-order thought - not because it is a result of blindly adding the 'I think' (why not a different phrase, after all?). Furthermore, the constant relatum S in the neural model of metarepresentation sketched above would effectively be the Cartesian theater and the physical relation would be presentation of a content in it. Hence at the vehicle level, we should think of metarepresentation and the object representation as two different but causally related neural states where the

former extracts and transforms some of the features coded by the latter - and this is a general description of neural representation *simpliciter*. To repeat, whether a representation identifiable with neural activation is *about* another representation or merely processes some information contained in it is a distinction at the content level: it transpires in what the neural representation is used for.

Furthermore, we should distinguish metarepresentation in the strong sense from metacognition. Metacognition can be defined as the capacity of a (control) system to access information about its epistemic states (e.g. the level of uncertainty of some feature discrimination) and to select an appropriate action based on that discrimination (e.g. not taking chances if the uncertainty is high). Importantly, as Proust (2003) points out, the metacognitive capacity may be only *procedural* - no semantic representation of the metacognized state is needed. For example, in the case of uncertainty monitoring the control system (the brain) may employ a simple heuristic of relying on the activation strength of the feature discrimination as a proxy for its uncertainty. Since activation strength is a property of the *vehicles* of the lower-order content, such metacognitive state would not count as metarepresentation in the strong sense of representing a lower-order representation *qua* representation (although it would count as metarepresentation in the weak sense of being simply *about* the lower-order representation).[15]

The consequence of this discussion is that any theory where the concept of metarepresentation plays a key explanatory role have to either show what it is about neural tokens of metarepresentations that make them crucial in the objective explanation of consciousness, or concede that the concept of metarepresentation is important only at the content level.

[15]Procedural metacognition may nevertheless be a precursor of full-fledged metarepresentation. It seems plausible that thanks to the process of representational redescription (Karmiloff-Smith (1992)) and acquisition of the theory of mind, we eventually learn to form representations that are about representations as such. In this simplified picture, the theory of mind provides the concepts necessary for such an explicit representation while the metacognitive process provides the subject matter of the metarepresentation - the content conceptualized in the metarepresentation.

3.3 Do we need neuroscience to explain the unity of consciousness?

Hurley (1998) builds on the content / vehicle distinction and raises an important point that vehicle-oriented accounts of consciousness "stand in a *token-explanatory* relationship to the mental, rather than a *type-explanatory* relationship."[16] That is, explanations referring to vehicles of conscious states (usually neural activations) explain at best why particular token of mental states occur, not why each token bears the content it does. Now, given that 1) contents are types, 2) the unity of consciousness is manifested at the level of content (being conscious of mutliple things at a time), and 3) type physicalism is false,[17] it would follow that there is very little prospect of neuroscience shedding some light on the unity of consciousness.

To reject the informal argument above, I will take two steps. The first consists in showing that no account of the unity of consciousness can be based solely on content and consequently that while 2) is valid, the unity must be accounted for at the level of vehicles of content.

The second consists in realizing that the argument is valid only if a) we limit the import of neuroscience to causal explanations only, and b) causal explanation of mental states is all there is to scientific explanation of consciousness. Of course, neuroscience usually seeks to give a causal account of some mental phenomenon - this, after all, is the golden standard of scientific explanation. However, it is not limited to causal accounts only (see below). Next, even if we reject type physicalism, there still may be aspects of consciousness (independent of content) for which type identity with neural processes holds. Let me elaborate on these points.

[16](Hurley, 1998, p. 28)

[17]I assume the reader is somewhat familiar with the general discussion of anomalous monism and functionalism. Reviewing arguments against type-physicalism is beyond the scope of this work.

3.3.1 Hurley and the need for an objective account of the unity

Hurley (1998) argues at length what an account of the unity of consciousness ought to be like. First, she asks whether the unity of consciousness can be accounted for solely in terms of subjective, personal-level contents:

> Can the needed work of determining the unity or separateness at a time of conscious states be done solely by subjectively available resources, by resources internal to the contents of consciousness as traditionally conceived? Or must the work be done by something outside the subjective contents of consciousness, something objective - such as the objective identitites of persons, or bodies, or spatiotemporal locations, or neurobiological characterstics of the brain, or some subpersonal property? (Hurley, 1998, p. 100)

She convicingly argues for the latter: the unity requires an objective account. An example of a purely subjective account would be one according to which the unity is constituted by having a thought *stating* the unity of two component thoughts, as in "I am thinking that p and q." The idea behind subjective accounts of the unity is that it is constituted by self-consciousness in the form of a thought that is *about* mental states (the unity would be conferred, as it were, by a representation of the unity). This, by the way, would be a naïve reading of Kant's introduction of the transcendental unity of apperception by the 'I think' possibly accompanying every conscious representation (see the next chapter for a detailed explanation). Hurley puts forward a so-called just-more-content argument against the possibility of a purely subjective account of the unity. Its point can be summed up in the Kantian lesson that consciousness of unity (i.e. representation of A, B, ... Z as together) is not the unity of consciousness, which it presupposes. The just-more-content argument can be roughly summarized as follows.

Assume there are conscious contents p and q. From (1) It is thought: p, and (2) It is thought: q, it does not follow (3) It is thought: p and q.[18] The same argument applies even if the intentional object is a higher-order thought: from

[18]The 'It is thought' clause is to signify intentional directedness and its formulation reflects the allegedly Lichtenberg's objection to Descartes that in his methodical skepticism he was not entitled to claim 'I am thinking', only to claim 'thinking is going on'. However, Hurley's

(4) It is thought: I am thinking that p, and (5) It is thought: I am thinking that q, it does not follow (6) It is thought: I am thinking that p and q, even if (5) and (6) were thought at the same time. The reason is that nothing in the content of (4) and (5) can warrant that the identity of the 'I' in the two thoughts is known. Even if we replace the 'I' with a definite description (e.g. 'person X, born on Y, with physical features of Z'), it would still be possible that two conscious subjects would entertain the same thought, conceiving of themselves using the same definite description. Hurley thus concludes that adding *just more content* to the states whose unity we need to explain will always be open to the objection that thoughts with the same contents could be entertained by different subjects.

An implication of the argument is that an account of the unity of consciousness must involve things or relations outside those that are subjectively accessible. Hurley spends a lot of time considering the possibility of accounting for the unity in terms of coherence among contents (that is, contents are unified in one consciousness if they are coherent) to reach the rather obvious conclusion that coherence cannot *constitute* the unity because it is in principle possible that different subjects entertain thoughts that are coherent together as a whole. While coherence is a necessary condition for the unity of consciousness in the normative sense described at the beginning of this chapter, to fully account for it we need to find what constitutes the unity at the subpersonal level of vehicles of conscious contents. Hurley calls such an account 'objective', meaning that it must explain the unity in terms of other things than just contents of experience. In this sense, Kant's account of the unity in terms of spontaneity and transcendental consciousness of the acts of synthesis is objective because these features of the mind are not experienced.

3.3.2 Science and type-explanatory accounts of the mental

Before we turn to the question to what extent the rejection of type-physicalism thwarts the project of finding a scientific account of the unity, it should be noted that science often provides type-explanatory accounts of the mental as well. For

just-more-content argument holds even if we grant that the proper way to describe the implicit sense of intentional directedness is 'I am thinking'.

example, evolutionary and cognitive psychology often explain behavior in terms of hypothesized contentful states (beliefs, attitudes, preferences, intentions, etc.). Discussions of the theory of mind, metacognition, introspection, and cognitive development are highly relevant to consciousness studies and they are mostly type-explanatory. Theoretical neuroscience can also shed some type-explanatory light on consciousness. For example, the fact that most of the time our conscious contents comply with the coherence norm could be partly explained by theories that show how learning at the neural level (neuroplasticity and synaptic weighting) leads to an adaptive wiring whereby tokens of inconsistent types will inhibit each other.[19] Models and theories that in their explanation of some process refer to contents presuppose adopting intentional stance to the analyzed entity (Dennett (1989)). Something is interpreted as having a perceptual content, e.g. seeing a rock, if it can differentially act upon that content, e.g. examine the rock for signs of extraterrestrial life. The more vague the idea about some entity's intentions, the more uncertainty about what representational contents to ascribe to it.[20]

Now, explanations from the intentional stance are teleogical, they presuppose a final cause. Neuroscientific theories which in their causal explanation of mental contents and faculties justify their explanatory neural principles in terms of their adaptive purpose are therefore more convincing, other things being equal, than those which merely describe the principles of neural organization (which would suggest their contingency).

3.3.3 Content and type-token distinction

So, how seriously is the prospect of neuroscience to explain consciousness thwarted by rejection of type-physicalism? The most important reasons for re-

[19]Neural darwinism and predictive coding are examples of such theories.

[20]Unless we already have a theory of mind of that entity. In case of ascribing perceptual contents to human beings, we can rely on our folk theory of perception of which we know that it applies quite universally to all human beings. Once we turn to animals, things get more complicated. Researchers studying animal cognition have to take extra measures, compared to human subjects, to ensure that the intentions of the experimental subjects are clear beyond reasonable doubt. This often involves a long time of operant conditioning prior to the actual experiment to the effect that the animal wants to perform a task in order to get a reward.

jecting type-physicalism that are relevant to our question is multiple-realizability and holism of mental states attribution. Insofar as mental states are defined by their functional content, they are in principle multiply realizable, thus types of mental states may not be identical to types of physical states.[21] Regarding holism, having a particular mental state is not a fact isolated from other facts about what other mental states the subject can be in. For example, perception of a rectangular shape depends on subject's discriminative capacity with respect to other possible shapes. In contrast, identifying a physical type is independent of other physical types there are, although the functions these types serve may depend on there being other physical types. For example, that diamonds are constituted by carbon atoms placed in a particular lattice is independent of there being other lattices in which carbon atoms can be organized.

How does this relate specifically to the possibility of scientific explanation of the unity of consciousness? Let's first take the integration aspect of the unity. Neuroscience would need to explain in virtue of what are mental representations A, B, C united in an integrated representation [A+B+C]. In order to do this, it would need to be able to identify the component representations independently of the integrated information and vice versa. Ideally, it would then identify the causal mechanism X thanks to which if A, B, C, and X occur, [A+B+C] occurs. This brings up many complications. First, the occurrence of [A+B+C], in contrast to joint occurrence of A, B and C, may be ascertained only if the subject makes a practical inference that requires holding the component representations in one consciousness. For example, that a subject is conscious of both a child running in the street and a car driving there can be ascertained if the subject makes an action to prevent the child being run over. First thing to note is that there are many such inferences than can stand as evidence for the occurrence of integrated representation. Moreover, the situation here is assymetrical: the absence of an action requiring synchronic consciousness of A, B and C does not by

[21] Accepting that the relation of the mental to the physical is that of supervenience may be ontologically correct but, unlike rejection of type-physicalism, it has no implications for empirical research.

itself imply that the subject is not conscious of them.[22]This complicates the issue because neural correlates of thus identified integrated representation [A+B+C] will always be confounded by the "output" side of the practical inference. But suppose that the neuroscientist varies experimental conditions so that subjects make various inferences, all necessitating consciousness of [A+B+C], and that she would consequently identify the integration mechanism with that which remains invariant across all conditions (for the experimental manipulation would screen out the output part). Then still, given the assymetry between integrated representation and its behavioral manifestation, we could only hope to claim that the discovered mechanism is sufficient for the integration, not necessary.

Perhaps, the argument above should be interpreted as a *reductio ad absurdum* to the conclusion that it does not make sense to detach conscious content from its practical manifestation, in which case we would need to deny that there is an integrated representation unless this is necessitated by an action the subject makes.[23] This, however, leads to the sort difficulties related to behaviorism. To

[22]That is, unless we go full behaviorist about mental states. Although behaviorism has lost its appeal, it brings to attention the fact that mental representations are theoretical entities that play a prominent role in explanations in cognitive science. There are at least two reasons for assuming conscious representations without their overt behavioral manifestation. First, thanks to our theory of mind we are prone to attribute to people epistemic and perceptual states solely on the ground of their location, state of their sense, or history. Second, the representational framework offers successful explanations, and because representations in this framework are conceived of as independent of the actions they lead to, it makes sense to assume consciousness of A, B and C without any apparent action that would require such consciousness.

[23]Dennett, to name just one, would certainly agree with this conclusion, while Chalmers would probably deny it on the grounds of the conceptual possibility of zombies. If zombies show that there could be conscious-like behavior without consciousness, then the possibility of consciousness without its behavioral manifestation should also be possible. This is a consequence of the zombies argument's conclusion about the conceptual independence of consciousness and behaviour: the independence relation is symmetrical. Although I don't accept the zombie argument, extreme cases such as patients with the locked-in syndrom (patients are completely paralyzed while their brain functions normally) provide compelling evidence of consciousness without behavioral manifestation. However, we can accomodate this with the Dennettian view by arguing that the extreme cases of consciousness with no manifestation are parasitic on the manifestation of consciousness in standard conditions. In other words, the paralyzed person is

find a way out of this conundrum we need to reconsider what kind of experimental evidence could be used to argue for some mechanism of integration.

The neuroscience research procedure described above in the example of a neuroscientist searching for X, the integration mechanism, is typical for fMRI studies which rely on the common method of cognitive subtraction. The purpose of this method is to study the neural mechanism behind a chosen process by contrasting neural activations under two tasks which supposedly differ only in that one involves the process and the other does not, other things being as similar as possible. For instance, face recognition process is studied by using normal faces as the target stimulus and scrambled faces (oval shapes filled with randomly placed facial elements, to control for visual complexity) as the control condition. Regions of the brain that show statistically significant difference in activation during this contrast are then considered to be implicated in the studied process. There are many questionable assumptions underlying this method, but it suffices to say that it is more suited to account for low-level, highly modular processes.[24] Another approach would be to hypothesize a neural integration mechanism and then manipulate it experimentally. If it happened that interference with this mechanism results in a failure to integrate information and therefore to make a practical inference, while preserving the component representations, it would be evidence that the mechanism is *necessary* for integration. Again, a similar problem arises with identification of unconscious representations A, B, C, for it would need to be shown that while tokening of [A+B+C] was disabled by inhibiting the integration mechanism, A, B, and C were still represented unconsciously (or co-consciously with other representations).[25]

---

conscious only thanks to the previously learned (and later internalized) manifest expressions of mental states.

[24]One of the assumptions is that the target process does not influence, e.g. by feedback connections, the activity corresponding to processes common to both the target task and the control task. For further discussion of the cognitive subtraction method, see for example Price and Friston (1997).

[25]If it is not possible to ascertain that A, B, and C are tokened, the function of the hypothesized mechanism cannot be interpreted solely as integration, for an alternative interpretation could be that the mechanism has some role in the tokening of the individual representations as well.

The main advantage of the causal research (in contrast to the essentially correlational research sketched above) is that it can, in principle, provide more substantiated claims about sufficiency or necessity of a hypothesized mechanism of integration. However, it does not obviate the methodological problems of distinguishing integrated vs non-integrated representations.

3.3.4 Self-consciousness and content-vehicle distinction

Having discussed methodological problems related to neuroscientific research of integration unity, we shall discuss now the prospect of neuroscience to provide an explanation of self-consciousness, specifically of the capacity for self-reference without identification. To study this empirically, we would ideally need a single dissociation between a conscious mental state and awareness of having that mental state. Such dissociation seems to occur in the case of the implanted thoughts delusion (one of the symptoms of schizophrenia) in which a person reports thoughts while denying they are hers. The afflicted person is aware of the thought, she just fails to attribute the thought to herself, non-inferentially. One could then argue that this is not the sense of self-consciousness that is interesting - don't we, after all, want to understand how awareness as such, i.e. that which even the delusional person has, is possible? Yes, but self-consciousness in this broad sense is coextensive with consciousness *simpliciter*. Since we cannot conceive of a dissociation between consciousness and self-consciousness in this wider sense, the relevant experimental contrast would be that between conscious and unconscious representations.

However, the implanted thoughts dissociation is not free of methodological difficulties either. Suppose that a neural mechanism is found such that manipulating it experimentally induces the implanted thought experience. Suppose further that researchers could induce this experience for any occurring target thought. It could be argued that the discovered mechanism is implicated in the positive attribution of the thought to some unknown other, not in the absence of self-identification. In other words, the mechanism could stand for the process of *interpreting* an occurrent thought that lacks the 'my' tag, so to speak, as someone else's thought - not for the process responsible for the absence of the 'my' tag

as such. To differentiate the two interpretations, we would need either a way to establish the absence of self-attribution (not just the delusional attribution to someone else) or a way to subtract the positive attribution itself. The first would likely lead to a situation in which the conscious/self-conscious distinction would again collapse into the conscious/unconscious distinction, and the second seems to bring more methodological complications than it solves.

Besides methodological problems, self-consciousness poses the conceptual problem of how to understand the relation between the object mental state and the reflective, higher-order state. Hurley cautions that attempts to account for self-consciousness naturally but erroneously lead to conflation of the content/vehicle distinction. Since we hold that an account of self-consciousness is a crucial part of explaining the unity of consciousness, we shall look at Hurley's argument more closely.

She starts by noting that it is common to identify conscious states with vehicles of the conscious contents (e.g. neural activations). We also assume that the content of a conscious state is its essential property. As a consequence, the vehicle is assumed to carry its conscious content essentially. The problem arises when we make a further assumption that the content a vehicle carries is its *intrinsic* property, e.g. that a particular pattern of neural activation carries that content independently of other possible mental states or interactions with the world. In other words, we tend to overlook the possibility that a vehicle carrying a specific content could be a relational property (relational to other vehicles/states) and still essential. Hurley's diagnosis of the tendency to think that contents must be intrinsic to their vehicles is that if we allowed them to be relational properties, there would no principled reason to hold that the relations must be confined to one's head. And we have a strong intuition that vehicles of conscious contents must fall inside the salient boundary that is our skull:

> We assume that intrinsic properties of vehicles must fall inside the boundary [one's head or body], because we assume vehicles must. [...] The boundary assumption may also explain the slide from essential to intrinsic properties of vehicles, where conscious content is in question. This slide overlooks the possibility that the essential properties

> of vehicles of conscious content are relational. After all, computational states are functional states, which can be realized in different ways. Their role in relation to other states is essential to them, not their intrinsic properties. But if vehicles of consciousness go relational inside the head, why couldn't they in principle go relational outside the head? That would violate the boundary assumption: the assumed boundary would no longer capture intrinsic properties of what carries conscious content. (Hurley, 1998, p. 35)

Hurley herself is in favor of vehicle externalism about consciousness, i.e. the idea that vehicles of consciousness include also things outside the organism. The important point, however, is to recognize that vehicle internalism should not lead us to assume that the content is fixed by intrinsic properties of the vehicle. Content can be an essential *relational* property of vehicles even if these are thought to be located only within the boundary of one's body.[26]

Now, self-consciousness perpetrates the slide to the idea that content must be intrinsic to vehicles because self-consciousness is thought to be self-evident: it directly reflects the content of the object conscious state. And if both the object conscious state and the reflective conscious state are identified with vehicles, how could the latter access the content of the former if the content were not its intrinsic property? Here the mistake consists of thinking that the subject can reflect on the contents of her conscious states in virtue of one conscious state having access to the content of another conscious state. But what stands in the access relation to contents is not another conscious state (identified with vehicles), it is the subject (whatever that turns out to be in naturalistic terms). As Hurley points out, to think otherwise is to enter the infinite regress of explaining subject access in terms of smaller subjects accessing some content. The relation between the vehicles of the higher-order state and the object state need not be any more mysterious than the relation between, say, a low-level and high-level representation of a visual

[26]See Clark (2009) for an argument for internalism about vehicles of conscious states that makes room for externalism about their content. Basically, the argument suggests that human body acts as a low-pass filter for information and hence that the information transfer which needs to occur at the right speed to yield consciousness can occur only within the body, in the nervous system.

scene. That is, the causal/functional relation between them must be such that, considering their relations to other functional states, it is correct to say that the content borne by one is a metarepresentation of the content borne by the other.

3.4 Summary

This section have showed that the unity of consciousness spans across various conceptual and ontological levels. It has both a normative and an objective aspect. The coherence criterion dictates that contents unified in one consciousness should not be in obvious contradiction. This applies not only in the sense that when we ascribe conscious contents to a person, we do so using the coherence criterion, but also in the sense that an agent will *usually* entertain coherent contents because failure to do so would result in maladaptive behavior. Next, ascriptions of conscious contents rely, among other things, on the assumption that the agent is rational. This assumption is crucial for determining whether the agent integrated relevant information in her consciousness. Without it, we could not hope to dissociate the process of integration from the process of verbal reports of conscious contents.

Next we noted that the discussion of (the unity of) consciousness needs to address both the level of conscious contents and vehicles of conscious contents. Rejection of type-physicalism complicates the prospect of explaining the unity. Although neuroscientific theories can provide type-explanations of why and how the brain realizes some conscious representations (e.g. by evolutionary and computational reasoning), the actual neural mechanism responsible for integrating information can only concern tokens of mental states (for causal relation stand among tokens of mental states).

We then discussed few methodological issues specific to neuroscience, the main conclusion being that although there are good reasons to talk about neural representations, it is very difficult to identify a token (neural representation) of the integrated conscious states [A+B+C], and, in consequence, the causal mechanism responsible for the integration. The remaining option seems to hypothesize a neural mechanism explaining how distinct representations hang together (the binding problem). Here we cautioned that not every common feature that the

conscious representations share at the neural level is necessarily implicated in consciousness. Manipulating the proposed mechanism and observing the effects on agent's practical inference is here the only option to ascertain its constitutive role for consciousness.

4 Kant on the unity of consciousness

4.1 Preliminary remarks

As Brook (1997) states, most of what Kant says about the mind comes from *The Critique of Pure Reason*.[1] Although Kant's project is primarily epistemological, he puts forward an account of consciousness that explicitly treats its unity as a fundamental feature. The goal of Kant's epistemological project is to show how it is possible that the world conforms to our knowledge. Or, in Kantian jargon, how synthetic a priori judgments are possible. Part of the argument consists of analysing what the mind must be like to be able to represent objects, which is considered to be a self-evident premise.[2] The B deduction emphasizes that representing objects requires unified consciousness, the synthetic unity of apperception. Kant's insights into how the mind is organized can thus be viewed as a result of exploring preconditions of unified consciousness. His transcendental psychology is thus particularly important for our project.

The requirement of unified consciousness is perhaps clearer for empirical self-consciousness, i.e. self-attribution of a mental state. I could not recognize myself as the subject of a mental state if the manifold of sensibility, from which the conscious mental state is synthetised, was not synthetised in *one* consciousness. To put it in words that do not appeal explicitly to Kant's terms: in order to realize that I am representing an object, that representing (as a process of integration of information from various sources) must take place in the same consciousness.[3]

[1] Unless stated otherwise, all following citations of Kant's CPR are taken from the Kemp Smith's translation. All references to CPR are in the standard pagination of the 1st (A) and 2nd (B) editions.

[2] Cf. (Kitcher, 1993, chpt. 3)

[3] (Kitcher, 1993, p. 95) argues against Strawson (1966) that apperception should not be regarded "as the unargued first premise of the transcendental deduction", because she takes

Following this line of thought, Rosenberg (2005) offers a comprehensible interpretation of Kant's argument starting with the fact of empirical self-consciousness:[4]

1. We are able to think of ourselves as the unitary subjects of our mental states.

2. Without a unitary comprehensive experience of the world, we would not be able to think of ourselves as the unitary subjects of our mental states.

3. Synthetic a priori judgments allow us to form a unitary comprehensive experience of the world.

4. So, synthetic a priori judgments are possible.

Now, whether Kant's argument is succesful or not has been a matter of dispute. Insofar as we are concerned only with the unity of consciousness, we can ignore the question of epistemic legitimacy of synthetic a priori judgments and focus solely on arguments for statements 1-3. To put it differently, we may focus on Kant's transcendental psychology, while leaving aside the metaphysics of transcendental idealism. As Kitcher (1993) remarks, many scholars have little respect for Kant's transcendental psychology since it is *prima facie* inconsistent with the tenets of his own critical philosophy. How can we be sure of the truth of the conclusions of transcendental psychology about the mind's functioning if the transcendental features of the mind are not, by definition, presented to us in experience?[5] Thus Strawson disregards the doctrine of transcendental psychology and interprets the

Kant to be aware of Hume's skepticism about mental unity - skepticism that Kant wanted to show to be incoherent, as she holds. Her exposition of Kant's implicit argument against Hume goes along the line suggested above.

It is interesting to note that the argument (based on the necessity of synthetising a flux of information) assumes that information, at the level of sensibility (that to which we are passive), *is* a constantly changing flux. What ground do we have to assume this, apart from our *empirical* theory of sense perception? By definition, we do not experience affections at the level of sensibility. Thus the assumption is not so innoucuous. Be as it may, the debate about the actual starting premise of TD is less important for our purposes than the actual characterization of the unity of consciousness.

[4]For the unabbreviated interpretation of Kant's argument, see (Rosenberg, 2005, p. 58).

[5]For example: "The theory of synthesis, like any essay in transcendental psychology, is exposed to the *ad hominem* objection that we can claim no empirical knowledge of its truth;

arguments in the TD concerning the necessary features of the mind as showing what is entailed in our concept of (conscious) experience.

Kitcher (1993) and Brook (1997), on the other hand, defend a more psychological interpretation of the findings in the TD. Their line of defense is based on interpreting Kant as holding a functionalist view of the mind and consequently interpreting transcendental psychology as revealing what functions any mind must realize. Finding how these functions actually are realized is then a matter of empirical research, and necessarily so due to multiple realizability entailed in the functionalist view of the mind. This is also the reason why they explicitly mention the import of Kant's work (and their own interpretation) for cognitive science.

I lean towards this view as well. The task of cognitive neuroscience can be understood as reverse-engineering the brain - for that we first need to know what functions the brain realizes and only then can we try uncover their neural implementation. Now, Kitcher is right that we cannot know, *by transcendental reasoning only*, that the brain is causally responsible for the mind, or generally that the mind is realized materially.[6] The currently widely accepted view that brain processes constitute the mind is based on *empirical reasons*. Thus the investigation of how the brain could realize the syntheses necessary for the unity consciousness is clearly not congruous with Kant's critical philosophy. However, this does not mean that we could not take Kant's analysis of the necessary features of the mind and try to explicate them in the naturalist, as opposed to transcendental, framework.

4.2 The unitary subject of mental states

The general thesis that we are able to think of ourselves as the unitary subjects of our mental states follows from the fact (which is stated, not argued for) that we

for this would be to claim empirical knowledge of the occurrence of that which is held to be the antecedent condition of empirical knowledge." (Strawson, 1966, p. 32)

[6]Specifically, (Kitcher, 1993, p. 203) criticizes Searle for his argument that intentionality (which we obviously exhibit) must be a result of the brain's wetware because it cannot follow from mere instantiation of a computer program (as the Chinese room argument intends to show). She thinks Searle's argument has the same structure of the rational psychologist's argument for the immaterial soul that Kant criticizes in the paralogisms.

are sometimes empirically aware of our mental states. What Kant wants to get is the thesis of the necessary unity of consciousness (a.k.a. the transcendental unity of apperception). The thesis is best expressed in his famous dictum: "It must be possible for the 'I think' to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me." [B131-2] That representations *are something* to me follows simply from the fact that we are capable of awareness of our mental states. Importantly, Kant formulated the unity thesis in terms of potentiality: "it must be *possible* for the 'I think' to accompany…". Looking at it from the perspective of representations, a representation (qua act, i.e. a represent*ing*) does not need to be actually represented in order for us to be conscious of it, it only needs to be available for inner sense.[7] Hence, the extent of consciousness (conscious contents as opposed to the unconscious, using the modern distinction) is delimited by the availability of mental states for empirical self-consciousness, or - to use a contemporary term - a higher-order state. It is important to keep in mind that this does not mean that availability of mental states for empirical self-consciousness actually *constitutes* consciousness, because the former depends on the conditions of possibility of conscious experience, rather than vice versa.[8] An interesting question then is what it is about conscious experience that, in Kant's view, enables empirical self-consciousness and delineates it in precisely this way.

The unity thesis states that we must be aware of the many representations collectively as mine. Or, to put it differently, we must be aware that the representations belong to me *qua* their common subject. The status of the thesis is best thought of as a description of some minimal condition of what it takes to experience things. For Kant, any meaningful notion of experience must entail

[7] This remark is redundant at best when considered from Kant's perspective, because the idea of 'unconscious' representations would be meaningless for him. What contemporary cognitive science classifies, from its naturalist view of the mind, as unconscious representation would fall, in Kantian terminology, under the concept of unsynthetised manifold of sensibility.

[8] Clearly, self-consciousness is a special kind of consciousness, therefore explaining the latter by the former would beg the question. That is the reason why proponents of higher-order theories of consciousness must start with the general concept of representation and try to explain consciousness in terms of relations among representations.

this thesis.[9] As such, the thesis could be regarded as a mere explication of the meaning of (conscious) experience. To demonstrate that the thesis is substantial, let's consider the difficulties associated with Hume's skepticism about a single subject having various mental states.

Hume famously argued that when he looks within, he does not find any recognizable self on top of mental states themselves - the self of self-reflection does not have any perceptible qualities. And if we can't recognize the 'I' in 'I think X' and 'I think Y' as the same, how can we represent ourselves as the unitary experiencing subject? In Humean line of thinking, it is tempting to say that *we* simply make the mistake of attributing our mental states to a persistent being that is in fact just a succession of representations. Note, however, that in order to express this very idea, we cannot avoid conceiving of ourselves as unitary subjects - for to whom does the supposedly mistaken "we" refer?

We cannot but think of a perception of change as a succession of representations (or various synchronic representations, e.g. in looking at a visual scene) belonging to the same consciousness, for otherwise there would be no succession but rather a dispersed and unrelated multitude of representations, which, to quote Kant again, "would be nothing to me." And to anticipate Kant's argument, it is precisely the connectedness of representations (their synthesis) that allows for the unity of consciousness.

At the same time, however, we must not succumb to the natural but fallacious line of metaphysical reasoning employed by Descartes in his argument that the subject is a simple and immaterial substance. Kant mounts an argument against this in the paralogism of the soul. The fact that we apply the 'I think' in the transcendental unity of apperception without identifying the subject by any properties (self-reference without identification)[10] should not mislead us into thinking that we thereby acquire some special and immediate knowledge about

[9] Cf. (Strawson, 1966, p. 25)

[10] "In attaching 'I' to our thoughts, we designate the subject only transcendentally ... without noting in it any quality whatsoever—in fact, without knowing anything of it either directly or by inference [A355]."

the underlying reality of the subject.[11] This is why Kant emphasizes that the 'I think' is the general form of consciousness, rather than a representation, and cannot yield any knowledge of the self as such.[12] The identity of the subject across many representations is given to us not through an intuition (which is what would make the use of the substance category legitimate) but by some alternative means. Kant does not explain clearly what the alternative is though. His way of putting it could be summarised as knowing by doing: "The mind could never think its identity in the manifoldness of its representation ... if it did not have before it its eyes the identity of its act whereby it subordinates all synthesis of apprehension ... to a transcendental unity." [A108] Unfortunately, at the key part of this formulation he resorts to a metaphorical expression ("...have before its eyes..") and consequently does not elucidate the character of transcendental self-consciousness any better than saying that it is not the category of unity.[13]

In the next two sections of this chapter I will try, with the substantial help of Kant's interpreters, to go beyond the metaphorical expression and clarify what underlies the unity of consciousness and the known identity of the subject across multiple representations. For now, let me summarise the most important characteristics of the unity of consciousness (the transcendental unity of apperception):

1. In the unity of consciousness I conceive of myself collectively as a subject of many representations.

2. It is a condition of possibility of empirical self-consciousness. That is, we could not form an explicit representation of our mental states if our consciousness was not unified.

3. This unity is not experienced. Rather, it is the form of (conscious) experience. The form is that of the 'I think' that must possibly accompany

[11] This point cannot be emphasized too much as the tendency to think of the transcendental subject in terms of substance has led many, including the present author, to a more or less obvious version of the homunculus fallacy.

[12] "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." [B157]

"[T]he I that I think is distinct from the I that it ... intuits ...; I am given to myself beyond that which is given in intuition. " [B155]

[13] Cf. [B131]

all conscious states. That is, all conscious states can possibly be ascribed to the transcendental subject. The term 'transcendental' here stands in contrast with 'real' or 'objective', i.e. something of which we can form intuitions. Another way of putting it might be that such subject is just virtual or thought (as opposed to experienced).

4. The unity of consciousness allows for self-reference without identification. I can ascribe some properties to myself (those of having some mental state) without recognizing myself as one object among many.

These are the basic features of the necessary unity of consciousness. We can now turn to Kant's arguments about what the mind must be able to do in order for the consciousness to be unified in this way.

4.3 Unitary experience of the world - transcendental apperception from a logical point of view

Looking for the conditions of possibility of unified consciousness is Kant's strategy in the famous Transcendental Deduction. Interpreting this argument is a notoriously difficult task. Fortunately, we do not need to make a critical assessment of the whole argument, we only need to consider its general idea. The idea is that the unity of consciousness requires that the various sensations be so connected as to yield a coherent experience of the objective world. To explicate the idea schematically, it will be instructive to follow the notation introduced by Rosenberg (2005). Suppose we have distinct representations whose content can be expressed by propositions 'I think X', 'I think Y' and 'I think Z'. This does not warrant the claim of the "analytic unity of apperception" which would have the propositional form 'I, who think X = I, who think Y = I, who think Z.'[14] In order to be conscious of the identity of the subject of these three representations, which is the requirement for transcendental self-consciousness as described in the previous section, the condition of the synthetic unity of apperception must be met. This condition can be expressed as having a representation the content of which is 'I

[14]Cf. [B133] and (Rosenberg, 2005, p. 57)

think [X + Y +Z]'.[15] This synthesis into [X + Y + Z] is the work of understanding applying concepts (of which the most general are the categories) and apriori forms of intuition. Thus the synthetic unity of apperception, viewed as connectedness of representations, is ontologically prior to empirical self-consciousness and it is prior in the order of explanation to transcendental self-consciousness.

The connectedness that Kant has in mind is not like a logical conjunction of separate items, but rather something like a conceptual structure which yields a coherent experience of the world. The epistemic glue, as Rosenberg puts it, by which representations are synthetised (i.e. the '+' in '[X+Y+Z]'), must be some overarching concept which, as a principle of unity, brings the representations together in virtue of their content.

This can further elucidated if we consider how Strawson (1966) tries to explicate the character of experience. He interprets Kant as thinking that anything that can count as experience must have the general form of an encounter with an object, where object is contrasted with subjective representation of it:

> To know something about an object, ..., is to know something that holds irrespective of the occurrence of any particular state of consciousness, irrespective of the occurrence of any particular experience of awareness of the object as falling under the general concept in question. (Strawson, 1966, p. 73)

In other words, in order for a representation of something to be an instance of (conscious) experience, the representation must entail that the thing is such and such independently of our way of representing it. Hence, this objective representation requires the possibility of conceptualizing the world as it is and as it only seems to us. Or, as Strawson puts it, the order of things as conceived objectively must be different from the order of our experiences of them. This, of course, does not imply that the objective representation yields knowledge of things as they are

[15] To avoid confusion, the 'I think' in the notation used does not express empirical self-consciousness, i.e. an actual self-ascription of a thought, but rather intentionality or directedness at an object that itself can become a represented object in a reflexive act. For a more detailed exposition of the schematic illustration of Kant's point, see (Rosenberg, 2005, pp. 117-125).

in themselves. Our knowledge of the world is still perspectival and constrained by our cognitive faculties. The point is only that it is the complex idea of an objective world in which our representations are synthetised - our conception of the world must be such as to potentially accommodate any of our representation and thereby synthetise them in one coherent experience. The objectivity makes sense only in contrast with awareness of how things merely seem to be, therefore self-consciousness in this minimal sense is necessary for any objective (in the Kantian sense explained above) representation of the world.

Now, Kant would claim that this objectivity is necessarily involved in any meaningful conception of experience. An important challenge to this view is thus considering a conception of experience that would be constituted solely of (disconnected) sense-data and would not make any pretense of objectivity. The argument against this challenge begins with noting that if we were to regard experience as mere succession of particular sense-data, then the unity of consciousness would be merely stipulated:

> [H]ow can we attach a sense to the notion of the single consciousness to which the successive "experiences" are supposed to belong? We seem to add nothing but a form of words to the hypothesis of a succession of essentially disconnected impressions by stipulating that they all *belong* to an identical consciousness. (Strawson, 1966, p. 100)

What is needed is that our experience involve subsumption of particularities under general concepts. Since concepts abstract from individual differences and capture that what is invariant across multiple encounters with objects the concept applies to, they allow for the distinction between the representing and the represented; a distinction which collapses into one thing in sense-datum theories. This brings Strawson to his formulation that "[a] series of experiences builds up a picture of an objective world in which the order and arrangement of the objects of which they are experiences must be conceived of as distinct from the order and arrangement of the experiences that form the series."[16] Importantly, a necessary condition for this is that the subject also recognizes itself as an object in the world, i.e. as a body. Only if I know myself as a part of the objective world can I make sense of my

[16](Strawson, 1966, p. 104)

experience as the result of being situated and having a particular perspective. If I were to know myself only as the transcendental subject, I could not "model" my experience as a perspectival route through the world. Embodiment and enaction, to use the modern terms, thus seem to be crucial for the possibility of self-consciousness.

4.4 Kant's syntheses - transcendental apperception from a psychological point of view

The previous exposition of transcendental apperception, following mostly Strawson's interpretation, could be viewed as mostly semantic analysis of what self-ascription of mental states must entail. Kitcher (1993) explicitly opposes this interpretation of apperception and argues for a psychological one. According to her, transcendental psychology is ultimately empirical in that it concerns the phenomenal (as opposed to noumenal) self and the cognitive tasks which the mind must perform in order to represent things.[17] Her psychological interpretation renders apperception independent of self-consciousness:

> On my account, the "unity of apperception" refers to the fact that cognitive states are connected to each other through syntheses required for cognition. "Apperception" does not indicate any awareness of a separate thing, a "self," or even that different cognitive states belong to a separate thing, a "self." Rather, they belong to the unity of apperception in being connected by syntheses to each other. (Kitcher, 1993, p. 105)

Thus Kitcher understands the unity of apperception as what we would now call "information integration" (albeit a very sophisticated one). While Strawson suggests that the possibility of empirical self-consciousness is a necessary outcome of meeting the demands for objective representation (briefly: we could not attribute objective existence to something unless we could draw the distinction between our

[17] "Given his own doctrine of noumena, and an exhaustive dichotomy, however, the thinking self must be phenomenal. Hence, transcendental psychology must be empirical, in this sense." (Kitcher, 1993, p. 22)

perspectival view of the thing and its features that are independent of that) and that the doctrine of synthesis can be ignored in this respect, Kitcher disagrees. She emphasizes that the doctrine of synthesis cannot be ignored in a coherent interpretation of Kant because judgments and intuitions represent only in virtue of their causal connections to other mental states. In this view, the doctrine of synthesis is *a* theory of how the mind integrates information to yield conscious representation.

Perhaps a more important reason to study the doctrine of synthesis is that Kant speaks several times of consciousness of the act of synthesis (as opposed to the standard form of consciousness, i.e. that of a content) and suggests that it is (also) a necessary condition for empirical self-consciousness:

> For the empirical consciousness, which accompanies different cognitive states, is in itself diverse and without relation to the identity of the subject. That relation comes about, not simply through my accompanying each cognitive state with consciousness, but only in so far as I conjoin the contents of one cognitive state with those of another, *and am conscious of the synthesis of them.* [B133, emphasis added]

Unfortunately, the word 'synthesis' has the same sort of ing/ed ambiguity as representation. It is not clear whether Kant meant that we need to be conscious of the synthesising activity itself or just of its result.[18] The second interpretation is certainly more parsimonious than the first in which Kant would be committed to hold that consciousness can be directed at two different things: at a represented content and at its own activity. Moreover, the first interpretation might be difficult to reconcile with the basic principle that we cannot know our mind *per se.* Nevertheless, Kitcher seems to opt for the first interpretation:

> No individual cognitive acts can reveal the unity of apperception. This unity only comes about through the syntheses that must be performed on cognitive states for cognition to be possible and that create a synthetic unity across the states. *Further, we can only recognize*

[18]The issue here is not due to translation. Kant's original expression is "...und mir der Synthesis derselben bewußt bin". Both Kemp Smith's and Guyer-Wood's translations keep the ambiguity by talking about consciousness of the synthesis.

that unity and represent it to ourselves by recognizing these syntheses.
(Kitcher, 1993, p. 108, emphasis added)

'Syntheses', in plural, arguably refer to the activity, not the product (for the three syntheses do not yield different intermediate representations - they are rather three different aspects of the synthetic representation).

Thus, on the second interpretation empirical self-consciousness requires that I be conscious of the product of synthesis of representations, not just the individual representations. This is the idea that the synthetic unity of apperception, schematically described as 'I think [X+Y+Z]' (see 4.3), must precede the analytic unity of apperception. Or, as Brook (1997) argues, that I need to be conscious of a global representation. On the first interpretation, preferred by Kitcher, I must be conscious not only of the product but also of my synthesising activity (of the spontaneity of reason, in Kant's terms). In either interpretation, we need to understand what the syntheses are supposed to do.

4.4.1 The threefold synthesis

The unitary comprehensive experience of the world is achieved, according to Kant, by a process of synthesis in which the parts (the manifold of sensibility) are connected so as to form a whole. Regarding the doctrine of synthesis, Strawson (1966) makes an important cautionary remark that "we can claim no empirical knowledge of its truth,"[19] because it is a result of transcendental psychology which, by definition, goes beyond possible experience. But insofar as transcendental refers to conditions of possibility of experience, we can regard the doctrine of synthesis as an inference to the best explanation.[20] Since the purpose of this limited exposition of Kant's view is to clarify the notion of the unity of consciousness, I suggest adopting a pragmatic stance: let's see what Kant thought of the unity of consciousness and of the way it is produced by the faculties of our minds without assuming (as Kant probably did) that it is the only way.[21]

[19](Strawson, 1966, p. 32)

[20]See Brook (1997) for the interpretation of transcendental philosophy as abductive reasoning.

[21]It could be argued that such a cherry-picking approach to Kant is wrong because in order for it to be meaningful, one would have to work with a substantially different conception of the mind that is being analyzed than that which Kant assumed. Specifically, one could argue

Kant distinguished three different syntheses whereby the mind reaches knowledge from intuitions. They are: 1) synthesis of apprehension in intuition, 2) synthesis of reproduction in imagination, 3) synthesis of recognition in a concept. These three syntheses can be more or less directly explicated in terms familiar in cognitive science.

Synthesis of apprehension in intuition locates percepts on a common spatiotemporal structure. While space, according to Kant, is the form of external sense, time is the form of inner sense. To put it in other words, space is the dimension in which we locate perceived external objects (and our bodily sensations), and time is the dimension in which we locate our mental states, including the acts of sensing. From the perspective of cognitive science and taking vision as an example, the obvious correlate of this synthesis would be the integration of information along the dorsal pathway ("where" pathway) in the popular two-streams model of vision. The dorsal pathway realizes a cascade of neural processing starting from a retinotopic map and local receptive fields at the bottom of the hierarchy and ending with an integrated representation of relative positions of objects and their movement. The neural representation of time is much less known, but it is safe to assume that the brain is somehow able discriminate which event happened first, given any two events to compare (unless the real temporal difference is below a certain threshold set by the physiological properties of neurons such as the refractory period). The goal of this spatiotemporal integration is for an agent to be able to represent spatiotemporal relations of perceived objects.

---

that Kant analyzes the mind from a phenomenological perspective and that, as a consequence, the syntheses he distinguishes may not have a direct correlate in some information processing occurring in the brain, which is what the naturalists about consciousness would assume. After all, since Kant wants to build a new ground for doing philosophy, and this involves some exercise in transcendental psychology, he starts off in a theoretical vacuum and therefore what he uncovers at the beginning must be self-evident and true. But this line of reasoning would be mistaken, as Kant himself points out on the case of Descartes's reasoning. The fact that the doctrine of synthesis is formed at the beginning, in a theoretical vacuum, does not guarantee that it is the only possible explanation for what Kant sought to explain, namely the possibility of a unitary experience of the world and our self-awareness. It might be hard to find and argue for alternative views but that does not mean they are not conceivable.

Synthesis of reproduction in imagination refers to the activity through which an object can be represented even without its presence in intuition.[22] Imagination thus "fills in" potentially perceivable parts of an object that are not represented in the current sensation but which are "implied" by the objective representation of it through understanding. So, looking at a box, for example, I perceive it as a 3D object and part of the representation is also its back side which is blocked from my line of sight. This representation, Kant would argue, still belongs to sensibility.[23] Rather than an abstract geometrical representation, it should be better understood as a representation of the object from different perspectives.

Finally, synthesis of recognition in a concept refers to recognizing something as an object of thought or perception and thereby yielding an intentional mental state. For Kant, this does not necessarily involve what we now call semantic representation (e.g. classifying something as, say, a chair) because it goes to such low level of cognitive representation where an object is recognized as something which we might not yet be able to classify or which does not have a semantic content. It is basically a matter of individuating the manifold of intuitions into intentional objects on which I can focus my attention. A tone, a number, or a strange physical object are all recognized in Kantian fundamental concepts - the categories.

4.4.2 The synthetic consciousness

These syntheses are three different aspects of information integration, rather than three different processes that occur in succession. The question now is whether consciousness is unified solely in virtue of this threefold synthesis. Kitcher understands it as follows:

> [A] unity of apperception is created when cognitive states are connected through syntheses; a state belongs to a given consciousness

[22] Cf. [B151]

[23] "*Imagination* is the faculty of representing in intuition an object that is *is not itself present.* Now since all our intuition is sensible, the imagination, owing to the subjective condition under which alone it can give to the concepts of understanding a corresponding intuition, belongs to *sensibility.*" [B151]

> if it can be synthesized with cognitive states already connected by synthesis. (Kitcher, 1993, p. 119)

In accord with her interpretation of the synthesis as information integration, Kitcher holds that mental representations are synthetised (combined, connected) in virtue of their content.[24] This appears to be quite uncontroversial in the psychological reading of the synthesis. However, it seems to imply what in cognitive science is known as the bottom-up view of the construction of a conscious representation, and misses Kant's main point that representations are shaped by our conceptual faculty. Similarly, Keller (2001) argues against Kitcher that in her view, "synthesis, which Kant insists is a spontaneous activity of the self, becomes a function of the causal dependence of the self on stimuli."[25]

Note that in Kitcher's account there is no mention of self-consciousness. Although the conditions of consciousness have been partly deduced from the fact that we are able to attribute mental states to ourselves, self-consciousness does not clearly follow from the picture. This is quite confusing given that Kant sometimes employs the term self-conciousness to stand for the transcendental unity of apperception,[26] which, in Kitcher's psychological reading, amounts to information integration. Thus some interpreters put forward a reading of transcendental apperception that does not entail any consciousness of the self (Brook, Kitcher) while others (Keller, Strawson) hold that self-consciousness is implied in it.

The disagreement, however, is not substantial. It stems from different readings of Kant's Transcendental Deduction. The psychological reading of Kitcher (and Brook), aimed at extracting Kant's view of the mind, naturally leads to an interpretation of the key concepts that could be accommodated to the theoretical

[24] "Synthetic connection is a relation of contentual connection. Synthetic products are contentually dependent on synthetic progenitors." (Kitcher, 1993, p. 117)

[25] (Keller, 2001, p. 42)

[26] Notably at [B132]: "The unity of this apperception I likewise entitle the transcendental unity of self-consciousness, in order to indicate the possibility of a priori knowledge arising from it. For the manifold representations, which are given in an intuition, would not be one and all my representations, if they did not all belong to one self-consciousness. As my representations (even if I am not conscious of them as such) they must conform to the condition under which alone they can stand together in one universal self-consciousness, because otherwise they would not all without exception belong to me."

vocabulary of cognitive science. In the empirical framework of cognitive science, theoretical terms (such as representation, or the possibility of self-consciousness) need to be relatable to possible empirical findings, i.e. it must be possible to adjudicate whether the theoretical term was used correctly or not on the basis of experimental evidence. The epistemological reading, on the other hand, allows for interpreting the key concepts as referring to *logical* (not psychological) aspects of our experience. Thus Keller's and Strawson's insistence that self-consciousness is implied in any objective representation because the latter can be thought only on condition of having the distinction between the subjective and objective order of things is compatible with Kitcher's claim that transcendental self-consciousness is just a figurative way of saying that "[c]ognitive states belong to the unity of apperception only because some faculty in whatever material or immaterial form in which those cognitive states are currently realized or preserved creates synthetic connections among them."[27]

Again, the transcendental self which is implied in the concept of transcendental self-consciousness must not be thought of as an object. Taken as a referring expression, it should be best understood as the general idea of a point of view or perspective that abstracts from all determinations of a particular subject (a point of view), except for its general characteristics of representing something. Speaking psychologically, it is not a construct, it only refers to the connectedness of representations which renders them conscious (and *ipso facto* perspectival). Importantly, despite her psychological reading Kitcher agrees with Strawson that Kant did not present sufficient conditions for (empirical) self-consciousness.[28]

She adds that this is not a serious drawback since Kant's transcendental psychology set out to explain mental unity, not personal unity. While diachronic personal unity might not be crucial for understanding consciousness, the sufficiency conditions for empirical self-consciousness certainly are. For it is the occassional reflection of our experience what makes us explicitly realize that we are conscious

[27](Kitcher, 1993, p. 123)

[28]"[T]he categories and synthetic connection are not sufficient for self-ascription. So there is no especially close connection between apperception, the categories, and self-ascription, despite the current popularity of understanding the deduction in relation to these issues." (Kitcher, 1993, p. 127)

beings and motivates studies like the present one. A more down-to-earth reason would be that information integration alone would not explain the difference in degree of consciousness that we hold exist between humans and other animals. Although this does not mean that information integration cannot be specified in a way that would not render consciousness promiscuous among sentient beings without alluding to self-consciousness, the prominent role of self-consciousness in many theoretical accounts of consciousness (Kant's included) suggests that every empirical theory should account for it.

4.5 Two kinds of self-awareness

As mentioned earlier, Kant often characterizes the transcendental unity of apperception as a kind of self-consciousness. In other places, he talks about empirical consciousness and emphasizes that the transcendental unity of consciousness is primary to and necessary for it. Given his characterization of empirical consciousness, it seems to be clear that he means awareness of one's own mental states. Interpreters thus usually distinguish between transcendental and empirical self-consciousness.

Empirical self-consciousness yields genuine experience by applying concepts to intuitions originating in inner sense. According to Kant, every experience is a result of understanding applying concepts to the manifold of sensible intuition. Obviously, we have concepts which we can apply to a special object in the world, our self as a person. When we think about our dispositions or current mental states, such as being irritable, excited, etc., we ascribe these properties to an object, our self as a person, which we implicitly conceive of as enduring over time and perhaps not necessarily physical. It follows from Kant's epistemology that empirical self-consciousness does not provide us with knowledge of the self as a thing in itself, only as it appears to us.

Kant's commentators largely agree on the interpretation of empirical self-consciousness. Transcendental self-consciousness, on the other hand, is a lot more complicated. It is not clear whether transcendental self-consciousness is just a different characterization of the transcendental unity of apperception or whether it refers to some related but conceptually independent faculty of the mind.

One thing that is clear is that transcendental self-consciousness is not to be considered as knowledge in the Kantian sense, namely bringing inuitions under concepts. As a condition of the possibility of experience itself, it does not involve application of concepts. These negative statements, together with many of Kant's unclear claims in the TD about the transcendental self-consciousness, invite various interpretations. As we saw earlier, Kitcher's interpretation is that only empirical self-consciousness deserves the name - transcendental self-consciousness would be a rather misleading term referring to the availability of conscious states to introspection, i.e. empirical self-consciousness, in virtue of belonging to the same system.[29] Strawson interprets transcendental self-consciousness as referring to that character of our experience which enables us to distinguish the objective and subjective order of things. Thus transcendental self-consciousness would be the aspect of the mind's conceptual system thanks to which it can conceive of things as independent of its representations of them. Empirical self-consciousness would then be a matter of application of this distinction in a particular case, e.g. realizing that a stick in water only appears to be bent.

4.5.1 Brook's account of the transcendental unity of apperception

One of Kant's interpreters who is a strong proponent of the relevance of Kant's transcendental psychology to cognitive science of consciousness is A. Brook. In his book *Kant and the Mind*, he puts forward an account of the unity in representational terms that supposedly better fits with the conceptual framework of cognitive science. He also provides a very detailed interpretation of transcendental self-consciousness. His label for the notion is 'apperceptive self-awareness' (ASA). I will use this label to distinguish his interpretation from others.

Brook also starts with quoting Kant's dictum that the 'I think' must be able to accompany all my representations and argues that in ASA we are not only aware of our self as the subject of a single representing but also as the subject of

[29]Cf. "'Apperception' does not indicate any awareness of a separate thing, a 'self,' or even that different cognitive states belong to a separate thing, a 'self.' Rather, they belong to the unity of apperception in being connected by syntheses to each other." (Kitcher, 1993, p. 105)

many different representings at the same time.[30] Thus, ASA cannot be thought of as a merely formal operation in which a higher-order representing is formed by adding 'I represent ... ' to the lower-order represented content X. We showed earlier that being aware of oneself as the common subject of many representings requires that these representings are unified or synthetized in one conscious state. It is customary to call this subject, in which the many representings are united, the transcendental self. Positing such a self does not, however, imply any ontological commitment to such a thing *independent of* the reason why it was posited, namely to help our imagination or understanding of what it means to have unified consciousness. To repeat the point emphasized in the third paralogism, this transcendental self is only the logical subject of our thoughts and cannot be given in intuition. Furthermore, the transcendental self (as opposed to the empirical self, or ego, personality) should not be conceived of as a thing, although the use of a noun-phrase invites that. It can be a process, a pattern of relations, or even a representation itself, as Brook (1997) suggests.[31]

To avoid the confusions of objectual talk about the transcendental self as much as possible, let me describe the explanatory role that it and ASA play. 1) To say that our experience (consciousness) is unified in the transcendental self is perhaps the most intuitive way to characterize what the unity of consciousness means. That is to say: if we conceive of consciousness metaphorically as a space containing some contents, the transcendental self would be the container. 2) ASA is non-conceptual because it does not *identify* the self as one thing rather than another (whereas concepts enable us to identify a thing by its properties and to recognize two things classified under the same concept as numerically distinct). When I am aware of myself as the subject of representations, I am not aware of one out of more possible situations. 3) ASA is grounded in the act of representing

[30] To anticipate my interpretation, this awareness should be understood as something that enables the explicit representation of myself having various representations rather than as a special mental relation to the object represent*ing*s.

[31] "The someone to whom a representation is represented is probably not just a formal place holder on the model of 'It' in 'It is raining' (...), but it could in principle turn out to be almost anything else, ... In particular, I urge that we not assume *ab initio* that the subject of representations has to be something radically unlike the representation it has." (Brook, 1997, p. 30)

itself, not in some further representation as in case of ESA. So every representation can give rise to ASA. In other words, to be aware of oneself as the transcendental subject of a representation, one does not need any other but this representation.[32]
4) To be aware of oneself as the common subject of many representations, one needs to unite the representations into one global representation (Brook's term) via transcendental apperception. That is, one needs to have a mental state of the form 'I think [X+Y+Z]'. [33]

4.5.2 Global representation

Global representation is the key concept in Brook's account of the unity:

> Unity of consciousness = df: (i) a single act of consciousness, which (ii) makes one aware of a number of representations and/or objects of representation in such a way that to be aware of any of this group is also to be aware of at least some others in the group and as a group. (Brook, 1997, p. 38)

[32]"Our standard way of becoming aware of an *act of representing* is quite different from the way we become aware of any *object* of a representation. We become aware of acts of representing not by receiving intuitions but by doing them: 'Synthesis...., as an act, ... is conscious to itself, even without sensibility' [B153]; 'This representation is an act of *spontaneity*, that is, it cannot be regarded as belonging to sensibility' [B132]." ... "Doing an act of representing is also what makes me aware of myself as the agent of that act, the subject of that representation. When I am aware of myself as the subject of a representation, I am aware of myself not as a represented object but by doing an act of representing." (Brook, 1997, p. 79) To anticipate, the importance of sense of agency and awareness of one's own activity in general is echoed in the predictive coding theory, see 6.2, 6.2.6 and Taylor (2012).

[33]Note that it is not implausible to suppose that intentional directedness (the 'I think' part in our notation) may not necessarily be part of the represented of the global representing, since it is always implied in every act of representing. What *is* needed, however, is that the subject is at the same time aware of these representations as distinct, yet his. For if the global representation were some new gestalt representation in which the subject could not individuate component representations, the form would be simply 'I think A' instead of 'I think [X + Y + Z]' and consequently one would not be aware of oneself as the common subject of X, Y and Z representations, but only of the logically implied subject of the gestalt A representation. This is another way of saying that it makes sense to speak of the unity of consciousness only if we hold that consciousness contains distinguishable representations.

Obviously, the most complicated task is to explain how (ii) is achieved. To translate the formulation of (ii) to language closer to contemporary psychology, we need to show how some kind of association among representations arises so that they are tied together in the global representation. To use the notation introduced earlier, we need to show what the '+' operator in 'I think [X + Y + Z]' amounts to, and also how the group of conscious representations is delineated and recognized as such (what forms '[. . .]').

In line with other interpreters, Brook also understands the connections among representations as provided by the synthetic activity of the mind. However, he adds a psychological reading to it, namely that the connections ensure that being aware of one representation makes one aware of other representations. In my understanding, in saying this he has in mind the idea that focusing one's attention on a representation brings to consciousness those representations with which it is strongly connected.[34]

Interestingly, when Brook tries to expand on the nature of synthesis that forms the global representation, he invites the possibility that it is not a single capacity but rather a result of interplay of different cognitive capacities which consequently require some kind of unity themselves:

> If awareness must be unified across a range of representations and can be unified only when objects of representations are linked by acts

[34]Of course, one could ascribe a more transcendental reading to this part of Brook's definition, i.e. something along the line that awareness of one representation entails awareness of a group of representations. For example, awareness of a duck entails awareness of its shape, orientation, spatial location etc. I don't think that is what Brook has in mind mainly because 1) he often brings up concepts from cognitive science to suggest which mechanism could realize the function. For example, in relation to awareness of a group of representations he invokes the concept of chunking that comes from cognitive psychology of memory. 2) The suggested transcendental reading would not cover the unity of consciousness in the case of complex experience in which we presumably have component representations that can possibly be parts of different global representations. What needs to be explained is how a particular representation, for example that of the computer screen I am seeing now, is unified with another particular representation, for example that of the words I am reading from the screen. These two representations are independent in the sense that each can be part of a global representation while the other is missing, yet in the current case they are unified in one consciousness.

> of synthesis, that should hold implications for the various sensible and cognitive abilities used to perform these acts of synthesis. In particular, one would expect that these abilities would have to be unified, too. Perceptual, linguistic, judgemental (identificatory and feature-placing), volitional, memory, and other competences are all used in forming and manipulating at least a great many of our representations. The kind of unity would be different - it would be integration of competences, not unity of consciousness - but it would seem that anything that could synthesize objects, especially global objects that connect representations and/or their objects to one another, would have to have a highly integrated control system. (Brook, 1997, p. 39)

How do, then, ASA, the unity of consciousness and global representation relate to each other? Brook suggests that, according to Kant, consciousness just *is* the global representation. Consciousness is then unified simply by virtue of being one representation, and it has simply only one subject common to all that is unified in the one representation. This, however, only defers the explanation of the unity to the explanation of how the global representation is realized. For if we assume that there is indeed a global representation (as something distinct from mere set of component representations) and that is equal to consciousness, it follows from mere conceptual analysis of 'representation' that it is unified (because it is unitary) and that it has only one subject (every representing is done by one subject). What we need to understand is how the global representation is formed and how its structure (i.e. the nature of connections among its component representations) enables explicit consciousness of oneself as the common subject of many representings.

5 Self-reference and self-awareness

Before I assess how well the neuroscientific theories explain the unity, it is desirable to know what philosophical accounts of the unity are like. I have spent a whole chapter on Kant's account due to its complexity. This chapter will supplement that account with few more recent contributions that aim specifically at self-awareness. I will first provide a working definition of the unity of consciousness. After, I will focus on the work of Shoemaker and Castaneda who analyzed the logic of self-reference. I will then move to Hurley's two-level interdependence model of consciousness and her account of perspectival self-awareness. Finally, I will argue that empirical self-awareness ought to be understood as a result of re-presentation of the mental state that is being reflected on, rather than as a higher-order state containing the reflected-on state as its proper part.

5.1 The unity of consciousness: a working definition

The unity of consciousness = synchronic* connectedness of representations in a global representation such that the representing subject is thereby transcendentally aware of being the representing subject. Several qualifications need to be made.

synchronic*: as described in 2.1, this aspect of the temporal unity of consciousness refers to the fact that 1) at every moment we are conscious of things that just passed and sometimes anticipate things to come; and 2) this short temporal extension of momentary consciousness is not represented, i.e. it is not a result of ascribing past and present mental contents to the same empirical self as we do in recollecting things from long-term memory. I believe that this point is rather obvious. We could as well simply say that our conscious state *is* momentary, without adding the synchronic* qualification,

if we did not understand the concept of moment on the mathematically abstract model of time as a line with moments as individual points without size (to use a topological concept, moments should rather be conceived of as open neighborhoods). To borrow Bergson's expression, momentary conscious states have duration (*durée*). Kant also refers to this kind of temporal unity (see his famous example of counting) and argues for transcendental synthesis as its necessary condition.

connectedness: following Kant's reasoning outlined in the previous section, connectedness refers to various representations being tied in what could be regarded as a single coherent global representation. Another way to put it would be that at any moment the various representations must be embedded in a common context[1] or framework, if you will. For example, right now the context of my unified consciousness is writing a chapter of my dissertation. I am conscious of visual percepts of words on the screen, but only marginally conscious of other items on the desk that are still in my visual field - until my focus wavers, in which case the context changes, for example to the context of describing my current phenomenology. My auditory percepts, insofar as I am conscious of them at all, are embedded in it as unwelcome distractions. The struggle for finding the right words is the most salient inner sensation; but should the context change, I would become acutely aware of fatigue or hunger. The context thus determines which representations are conscious and which are not (the unconscious ones may be either irrelevant, as demonstrated in the famous invisible gorilla experiment, or directly incompatible with the current context, as for example in case of perception of the impossible triangle or devil's fork. The current context would be the principle of the unity of consciousness at a given time, to employ Kant's phrase. This is not to say that the context is something above and beyond the representations and their connections - rather, it is a label for the way the particular representations are related to

[1]The term context is here used in the sense B. Baars uses it in his *Cognitive Theory of Consciousness*.

each other.[2] For Kant, the ultimate context is the (idea of objective) world and the categories are the most fundamental forms of ties or relations among representations.

representation: The concept of mental and neural representation is elaborated and defended in 3.2.

transcendental awareness: Transcendental self-consciousness refers to the fact that I know myself to be the common subject of all my conscious representations without identifying myself in virtue of recognizing some properties. Transcendental self-consciousness is a necessary condition for empirical self-consciousness - ascribing properties (e.g. mental contents) to oneself as an object. To put it differently, 1) it must be possible to make any conscious state (a representing) an object of a higher-order representation; the higher-order representations count as cases of empirical self-consciousness, and 2) the identity of the subject of both the higher-order and the lower-order representations must be known non-inferentially, that is without identification through recognizing some properties.

subject: Taken from the phenomenological point of view, the subject refers to what Kant would call the transcendental subject, the 'I' of apperception, which, to repeat, is not to be thought of as an object with knowable properties. In contrast, from the naturalistic point of view endorsed in the following analysis, the subject can only be that to which we ascribe consciousnes - that is, the organism (if we hold that consciousness is a biological phenomenon) or the cognitive system (if we hold a functionalist view of consciousness).

Conceptually, the explanandum consists of two areas: 1) integration or synthesis of representations, and 2) transcendental self-consciousness. Following Kant, I argued that these two areas are two different angles at which one can look at

[2]Dehaene and Naccache (2001) articulates a similar idea in his account of the global workspace theory. The concept of context also plays a similar explanatory role in the predictive coding theory.

the unity of consciousness rather than two independent aspects. I treat them separately mainly for the sake of clarity.

The promising candidates for explaining the integration part would be those that account for the structure of mutual support between the global representation and its component parts as described informally in Brook's definition in 4.5.2. As for the transcendental self-consciousness, the crucial thing is to explain the known identity of the subject across conscious mental states where the subject is not recognized by properties. Next subsection will clarify the latter point.

5.2 Shoemaker and Castaneda: the logic of 'I'

Some of Kant's insights about the logic of unified experience are mirrored in studies of semantics and pragmatics of the first person pronoun 'I'. It is useful to consider these in detail for two reasons: 1) they are free of Kant's technical vocabulary, 2) they analyze what we are able to express in language which is less contentious than what we are able to represent.

At the beginning of his seminal article "Self-reference and self-awareness", Shoemaker notes that although 'I' is a referring expression, some philosophers found its referring role puzzling or even denied it. In an attempt to resolve the puzzle, Shoemaker first echoes Wittgenstein's distinction between two different uses of 'I' - the object use and the subject use.[3] The object use of 'I' denotes an object in the world (a body) to which objective characteristic might be ascribed (e.g. 'I am bleeding.'), and the subject use appears in expressions of propositional attitudes or mental states in general (e.g. 'I think it will rain.', 'I see a red table.'). The important difference between them is that the latter is immune to "error through misidentification relative to the first-person pronoun."[4] It might happen that I misidentify the bleeding person (for example, the blood on my forearm might be of someone else), but I cannot mistake someone else's mental content for mine.[5] As Wittgenstein puts it, it would be nonsensical to question whether it is really me who have pains in response to the expression 'I have a

[3](Wittgenstein, 1958, pp. 66-67), citation owing to Shoemaker (1968).

[4](Shoemaker, 1968, p. 556)

[5]Although one can probably mistake one's own thought for someone else's, as in the case of the implanted thought delusion in schizophrenia. If this interpretation of the schizophrenic

tooth-ache.'[6] Although other indexicals, for example the demonstrative 'this', are also immue to error through misidentification (because the reference is fixed by speaker's intention to refer to *that* object she has in mind), there are two important differences. First, while tokens of 'this' may refer to different things each time they are employed, tokens of 'I' always refer to the speaker and his intention therefore does not determine their reference. Second, while 'this' may fail to refer to anything (in case of hallucinations, for example - unless one holds the internalist view of perception), 'I' cannot fail to refer because the utterance must always be made by someone.

Shoemaker then continues saying that these puzzling properties of the 'I' expression led many philosophers to deny that it refers at all, or that it denotes a subject having the ascribed mental contents. This, he thinks, is false. He agrees with Wittgenstein that the subject use of 'I' cannot be substituted by a description of a body, or by any other expression free of all indexicals. The argument is simple and worth recounting: assume that I wanted to substitute 'I' in 'I see a red table.' by the expression 'the living human body located at *xyz* at time *t*'. The substitution would be justified only if I knew that *I* am the

experience is correct, it hints at peculiar assymetry of the immunity to error through misidentification.

[6]Shoemaker uses a slightly different example that is interesting with regard to classical experiments manipulating the sense of agency. He says that while 'My arm is moving.' uses the pronoun as an object and is therefore liable to error through misidentification, the statement 'I am waving my arm' is not. However, experiments manipulating the sense of agency (see Wegner (2005), for example) seem to imply that one could be mistaken even about this. After all, the subject of the experiment would claim 'I am moving with the mouse cursor.' when in fact it is the confederate who is controlling the cursor. Shoemaker could reply, using a concept developed later in the article, that it does not refute the idea of immunity to error through misidentification because the expression 'I am moving the cursor.' should actually be properly explicated in terms of underlying P*-predicates as 'I think/feel I am moving the cursor.', where the latter token of 'I' is in the object use. In other words, the statement is false in the contrived experimental setting not because the subject misidentifies herself, but simply because she wrongly ascribes the cause of the movement to oneself. The mistake is thus the same as if 'I won a lottery.' was simply false because someone else did. The false sense of agency is an interesting case, however, because it seems to yield an implict, unmediated knowledge of one's causal influence and is therefore different from the case in which I identify myself as a body in the world. How this unmediated knowledge could arise is described in 6.2.3.

human being located at such and such place and time. But in order to know that *I* am that body, I would need to identify myself with the body by means of some properties that I ascribe both to myself and the body, hence I would need to have some previous knowledge of my properties that is not yet derived from identifying myself as the body. (This argument is essentially the same as Kant's argument that transcendental self-consciousness is prior and necessary for empirical self-consciousness.)

Similarly, Castaneda (1966) argues that 'I' is the only demonstrative that cannot be eliminated by means of a description. While 'this' can in principle be eliminated by a description that expresses the sense of intentional directedness in the act of referring by that demonstrative (e.g. 'this' = 'the red table I was pointing to at time t'), 'I' (in its subject use) cannot be eliminated this way. I could not identify myself by a description if I did not know the decsription is valid of me - and that I must know only in a non-descriptive or non-inferential way.

To resolve the puzzlement about the referent of the subject use of 'I,' Shoemaker argues for the possibility of knowledge of myself (my mental contents) without recognition. He shows that the reason why some deny that the 'I' refer to some self in the case of the subject use is because they think of self-awareness on the model of sense perception - that it involves "seeing", with inner sense, a self that has some property, e.g. thinking that *p*. In this model, the supposedly perceived self is rightly denied to be a real thing, as Hume famously argued. It is at best an abstract placeholder to which we attribute mental states. Shoemaker admits that it is often difficult to avoid thinking about self-awareness on the perceptual model because the grammar of our self-attributing expression is highly suggestive in this way. Again, his reasoning is similar to Wittgenstein's argument against the possibility of defining one's mental states ostensibly (and privately). Although he insists that the 'I' in the subject use is an expression referring to myself, he does not elaborate on what this self is. While it may not be an interpretation Shoemaker would endorse, I suggest that the 'I' in the subject use be understood as Kant's transcendental unity of apperception. The 'I' would thus refer to the unified conscious experience from a specific point of

view at a specific time.[7] I take this interpretation to be at least compatible with Shoemaker's analysis.

Shoemaker then asks the crucial question how it is possible that "there should be [psychological] predicates, or attributes, the self-ascription of which is immune to error through misidentification."[8] The fact that he moves from purely linguistic investigation of the subject use of 'I' to considerations of its underlying psychological conditions implies that he takes the 'I' and its use to be not just a manner of speech but a substantial part of language or, for that matter, any system of reference. Any language rich enough to be able to express what the world looks like *from a certain point of view* must include an expression with the function of the subject use of 'I'. This idea bears a telling similarity to Strawson's reading of Kant's argument for the necessary unity of apperception. As described earlier (section 4.5), Strawson argues that objective representation (i.e. representation of something as an object, independent of my view), of which we are capable, requires the ability to distinguish between the subjective and objective order of things, and this in turn requires that I be potentially conscious of my representations as mine thanks to a contrasting representation of the world as it objectively is (where the latter is achieved by synthesis using categories etc.).

Insofar as any meaningful conception of language involves not just impartial descriptions of states of affairs but interacting speakers recognizing their commitment to justify utterances (the Sellarsian view), unified consciousness seems to be implicated in it. Although the study of linguistic practice and entailments of the subject use of 'I' as well as of pragmatics of the intersubjective commitment to justification could offer more insights into the fundamental role of the unity of

[7] Another possible interpretation is that the subject use of 'I' has a place in language mainly to designate a person that is committing herself to a statement and hence that the use is primarily intersubjective. In that case, the reference of 'I' would be the speaker as a person, not a momentary unified conscious state. But even if the expression referred to a person rather than the momentary state of consciousness, the question would arise of how immunity to error through misidentification is achieved. And to answer that question, the interpretation that 'I' refers to the unified state of consciousness provides a better starting position than the interpretation that 'I' refers to a person because persons are complicated constructs that presuppose consciousness.

[8] (Shoemaker, 1968, p. 565)

consciousness, I shall focus on the psychological aspect. This, to repeat, is the question "how it is possible that there should be predicates the self-ascription of which is absolutely immune to error through misidentification."[9]

5.3 Hurley's two-level interdependence model

As discussed in section 3.3, Hurley (1998) recognizes that the unity of consciousness needs to be accounted for both at the personal level of conscious contents and the sub-personal level of vehicles of conscious contents.[10] Her work has a substantial negative part, where she analyzes many misconceptions about consciousness, and a shorter positive part where she puts forward her two-level interdependence model.

The negative part consists mainly in exposing as untenable the conflation between the input/output distinction on the one hand and the perception/action distinction on the other hand. While both distinctions are useful, they do not clearly map onto each other. Conflating these two distinctions yields a false sandwich model of consciousness as something that perception/input is *to* and action/output is *from*. To show that perception does not always map onto input and action onto output, she reviews many neuropsychological cases showing that perception may change as a result of a change in output while the input is held constant, and vice versa for intentions. Hurley's speculative conclusion is that

> "the personal-level contents of both perceptual experience and intentions can in general be functions of the subpersonal *relations between* input and output, such as the relations that hold within a complex dynamic feedback system. Then the contents of perceptual experience and of intention will be essentially interdependent. Even though perceptual and intentional contents are different functions of the relations between input and ouptut, changes in these relations should in

[9] (Shoemaker, 1968, pp. 565-6)

[10] Personal level denotes here any discourse that presupposes the existence of persons (agents, thinkers, speakers, etc.), e.g. talking about actions, perceptions, intentions, etc. Sub-personal level description of an organism uses terms that do not presuppose this personal unity.

general be expected to affect both perceptual and intentional content."
(Hurley, 1998, p. 339)

Consequently, Hurley argues that consciousness research should focus on relations between input and output (or perception and action) rather than on the disputed relations between input and (conscious) perception. Having described several cases of interdependence of perception and action, Hurley introduces the notion of *dynamic singularity* of causal flows centered on, but not bounded by, an organism, and argues that this concept is a promising subpersonal complement to the personal-level normative criterion of coherence. Coherence of conscious contents alone is not a sufficient condition for the unity of consciousness, for it is in principle possible that a set of coherent conscious states is tokened, at a time, by two or more subjects.

The notion of dynamic singularity is somewhat vague. The general idea is to cash out the unity of a subject (agent) in terms of a third-personal description of a dynamic system. An agent is something that acts and perceives in the interdependent way described above. If we described the system of causal relations between input and output, we could say that the agent is simply the center of these causal loops. In this reading, the boundary between an agent and its environment is not clearcut (it certainly does not need to be located at the boundary of one's body, let alone the skull, as Hurley likes to point out). To take a simple example of such a causal loop, consider movement of one's arm. The subjective experience of raising an arm depends not only on formation of the motor intention but also on the corresponding proprioceptive and visual feedback. The continuous proprioceptive feedback provides information, subpersonally, whether the movement is being carried out according to the motor plan and drives small adjustments on the way. Notably, the lack of proprioceptive feedback leads to uncontrolled movements (vision does not offer detailed enough feedback for appropriate adjustments).[11] Should this causal loop of efferent motor commands and afferent proprioceptive feedback be disrupted, the subject would not experi-

[11]See the neurological case of Ian Waterman to find out more about the pathologies associated with loss of proprioception. Waterman was unable to control his movements and it took him considerable time and effort to adapt to the loss by relying on the visual feedback only.

ence the movement as intentional or real. For example, if somebody pulled my hand up, I would get a prioprioceptive feedback indicating the position of my joints but I would get no feedback from muscle spindles indicating stretch in the muscles needed to generate that movement myself, nor would I have tokened the intention to move that way. Similarly, if only the motor command were formed and carried out without the corresponding feedback, the action could be perceived as only imagined.

In order to substantiate the enactivist claim about interdependence of perception and action, Hurley often refers to the ecological view, represented most famously by Gibson's seminal work *The Ecological Approach to Visual Perception.* Gibson argued against the traditional view of visual perception as passive sensory information processing resulting in some internal 3D model of the world, and argued instead that perception be conceived as an active process of extracting information by means of scanning the environment from various perspectives (the dependence on movement, and in action in general, is thus implied) and to various needs (what we perceive are affordances that are relative to our needs).[12] Now, to get a correct perception of the environment, the agent needs to distinguish changes in sensory input induced by its own movement from those that are driven by changes in the environment. Note that a change in sensory input, e.g. tactile perception of fingers sliding through a cat's fur, is ambiguous as to the cause of that sensation. Either the cat could have moved, or I could have stroked her. What enables me to distinguish between the two cases is the sense of agency (and proprioceptive feedback in particular). Thus the embodied agent needs to keep track of which changes in sensory input are self-generated (and less obviously: which perceptual constancy is maintained by a self-generated movement, such as when we watch an object passing by while turning our head). Presumably, this is achieved by feedback loops between sensory inputs and motor outputs. If the feedback loop is disrupted, it severely impairs the agents capacity for nav-

[12]"Ecological theorists of action emphasize the interdependence of invariants in varied circumstances, yielding perceptual constancy, while perception guides as invariant intention through varied circumstances, yielding action constancy." (Hurley, 1998, p. 431)

igating the world. The famous experiment by Held and Hein (1963)[13] showed that kittens in the passive condition, i.e. those that were exposed to changes in sensory input only through passive movement in the carousel, subsequently failed at simple motor actions such as paw placement and safe descent in the "visual cliff" task. Another example is the case of eye paralysis in which an attempt to change the location where the eyes foveate yields the experience of the world suddenly jumping sideways because the expected saccadic movement is not performed (eye muscles are paralysed). The explanation is that in order to yield a relatively stable percept, the brain has learned to discount changes in retinal input corresponding to eye movements.

This sense of self-generated changes in input, that each complex enough agent needs for its succesful navigation in the world, is a very basic form of self-consciousness. Hurley calls it perspectival self-consciousness:

> [H]aving a unified perspective involves keeping track of the relationships of interdependence between what is perceived and what is done, and hence awareness of your own agency. In this sense, perspective already involves self-consciousness. But the sense of self-consciousness that makes good this thought is closely tied to ordinary motor agency and to spatial perceptions, and need not involve conceptually structured thought or inferences. (Hurley, 1998, p. 141)

Having a perspective is a natural consequence of being an agent situated in the world, and it does not require cognitive self-consciousness (Hurley's term for the explicit kind of self-awareness typical of humans). An organism posseses perspectival self-consciousness if what it does systematically depends on what it perceives and vice versa. In terms of dynamic systems theory, perspectival self-consciousness is thought to be constituted by action-perception feedback loops centered around the organism.

[13] In the experiment, ten pairs of newborn kittens were placed in a carousel in such a way that one kitten could move more or less freely around (the active kitten) and a mechanism transferred the motion symmetrically to a gondola where the other, passive kitten was placed. Apart from the experimental condition, the kittens were held in a dark room. The passive kitten was thus exposed to the same pattern of changes in visual sensation, but unlike the active kitten, it could not associate these changes with its own motor output.

This notion of self-consciousness is very general and hence widely applicable - to the point that it is questionable whether it tells us anything about self-consciousness as we know it.[14] Putting this aside, although self-consciousness is held to emerge from a multitude of rather specific interactions with the world, each governed by a relatively independent feedback loop, what makes them all draw on self-consciousness is that these loops are centered around the organism and serve to hold some of the organism's state constant.[15]

Note that the outlined mechanism of perspectival self-consciousness is described at the subpersonal level. This has the advantage of complementing the objective account of the unity of consciousness by explaining when tokens of coherent conscious contents belong to a single consciousness. Tokens of coherent contents entertained by someone else are not causally dependent on my tokens. The specific causal mechanism that underlies the dependence between tokens (vehicles of conscious contents) is a subject matter of cognitive neuroscience.

Hurley's account is the most convincing when it comes to explaining what it takes to be an embodied, perspectival agent, and what processes underlie having a point of view. In respect to the unity of consciousness, her main contribution is the concept of perspectival self-consciousness that can be taken as a sub-personal account of the most basic form of self-awareness and hence subject unity. She does not propose a positive theory of how conscious contents are integrated, besides

[14]One could argue that the description of perspectival self-consciousness is either trivial or too wide. It is trivial to the extent that the concepts of perception and agency, both used in the description, already entail the notion of perspective in a sense that is not further enriched by the description. And if perception and agency were to be cashed out in terms of dynamic relations between a system and its environment, it is perhaps too wide; for there are many systems that influence their environment and are in turn influenced by it - plants, robots, etc. The second option is less problematic, I think. It is plausible that perspectiveness comes in variety of complexity. Although I don't see a reason to assume that there is a natural threshold of complexity at which something becomes self-conscious in a strong sense, others may want to hold that self-awareness does not come in degrees and therefore either argue in favor of a specific threshold or deny the idea that perspectiveness is exhausted by a description of the agent as a dynamical system.

[15]Notably those states that are life-preserving. This, by the way, is one of the motivating ideas of the predictive coding theory under its free-energy formulation - see section 6.2 for further details.

hinting at causal dependence between tokens of mental states as a necessary condition for them belonging to one consciousness. More importantly, she admits she does not have a theory of what makes a representation conscious.[16] The main lesson to be taken is thus her detailed account of how subjectivity can be constituted by causal loops of action and perception of various orbits, each centered in the organism. This account is quite open to further amendments congenial to the enactivist view of the mind, so we can hope to supplement the missing points with ideas from other theories of consciousness.

5.4 Empirical self-consciousness

Hurley's two-level interdependence model is a good account of perspectival self-consciousness - the sort of implicit awareness of one's own states and situatedness in the world that every agent must possess. Let us now turn to the question of what constitutes the explicit self-awareness that is characteristic of human consciousness.

In chapter 4 we noted that the unity of consciousness is a prerequisite for empirical self-consciousness. We also noted that empirical self-consciousness is contrasted by Kant and his interpreters with transcendental self-consciousness.

I will try to outline an account of the capacity to reflect on one's content of consciousness (and thereby of the subject unity of consciousness) from a naturalistic stance by describing, from an evolutionary point of view, a process of cognitive adaptation possibly leading to self-awareness. This strategy is often employed in the philosophy of cognitive science to explain how some phenomenon characteristic of an intentional agent can be explained in a bottom-up fashion.

First, however, we need to distinguish two kinds of access to one's content of consciousness that can both be interpreted as self-awareness: the intentional and cognitive access.

[16] When addressing this issue explicitly, she concedes the possibility that what is needed, besides the dynamic singularity she described, is that it is instantiated by a living, biological organism. See (Hurley, 1998, chpt. 4.7)

5.4.1 Intentional access

When self-consciousness is thought to be constitutively related to consciousness, the sense of self-consciousness theorists have in mind is a specific kind of access to content. In the oft-cited case of blindsight, the phenomenologically blind person is said to be unconscious of visual information because she cannot report on what is around her and she cannot act intentionally (spontaneously) on the information that is nevertheless represented in the brain (where this is manifested by the fact that the subject succeeds in cued guessing). Hurley (1998) makes a correct analysis of what kind of access is actually constitutive of consciousness: intentional access, i.e. the ability to spontaneously[17] act on the accessible information, given one's intentions. Importantly, intentional access differs from cognitive access (the ability to form a belief about having the content p), so it is possible that the intentional access, and therefore consciousness, does not require conceptual representation.[18] A corrolary is that insisting on *cognitive* access as constitutive of consciousness makes the concept of consciousness more stringent, applicable perhaps only to humans.

5.4.2 Cognitive access

From the naturalistic perspective, intentional access seems to fall within the scope of unity as integration. It is a matter of integrating perceptual contents with a hierarchy of goals and processes subserving practical inference. Nothing deeply self-conscious seems to be implied. More specifically, something can

[17] That an action is spontaneous means, in this case, that it is not prompted by an explicit order from a third party, but rather originates from the agent herself. Of course, spontaneity so defined is not clearcut and it could be objected that which actions count as spontaneous is relative to the model (theory of mind) used by an observer adopting the intentional stance. But insofar as attributions from the intentional stance are constrained by patterns of behavior, there will be a wide agreement on which action is self-caused and which action is caused externally, although the boundary is fuzzy and some borderline cases will be disputed. After all, if spontaneity of action could not be ascertained beyond reasonable doubt, our legal system would have great difficulty adjudicating court cases.

[18] As Hurley points out, we may tend to miss this because as long as we discuss consciousness in the case of persons with conceptual and cognitive abilities, intentional access comes almost always with cognitive access.

exhibit intentional access and integration of contents across domains without necessarily treating the represented contents *as* representations. Remember that our current *explanandum* is the non-inferential awareness of my conscious states as mine, which presupposes awareness of represented contents *qua* representations, as opposed to the case where representations are transparent to the subject (i.e. when we are "at the things themselves").

To be clear, I don't want to imply that whether we are aware of content p (the thing perceived or thought) or our representation of p (the mental state of perceiving or thinking p) is something that can be fixed solely from the first-person perspective, by subject's phenomenology, as it were. On the contrary, the distinction between awareness of p and awareness of the representation of p can be substantiated only by reasons independent of subject's phenomenology. What kind of reasons could it be? The kind of reasons that make us attribute mental states in general: abilities to act in such a way that a parsimonious intentional explanation of them will refer to awareness of one's own representations (as representations). I will now outline two such capacities, taking up the evolutionary bottom-up explanatory style typical of naturalistic accounts of mental capacities.

5.4.3 A short evolutionary story of the origin of self-consciousness

Let's have an organism O that is capable of complex interactions with the environment that include dispositions to behave in ways that are at least to some extent context-sensitive. For example, it would fear and flee from a freely wandering lion, but not flee from a lion in a cage. Context-sensitivity is relative to the scope of domains that we are willing to include in a Skinnerian explanation of behavior in terms of disposition acquired by operant conditioning before we yield to explanation in terms of consciousness integrating information to allow for novel action. As a rule of thumb, if O can react to a novel situation in a way that is both rational and novel (not previously rehearsed), we are justified in thinking that O can integrate information from various domains and attend selectively to the relevant information (which manifests in the fact that the action is *rational*).

According to the cognitive theory of consciousness, this constitutes the minimal criterion of consciousness.[19]

In virtue of the complex interaction with the environment, O has mental, contentful states. That is, O's behavior can be successfully predicted from the intentional stance and its patterns of interaction is what fixes the content of the attributed mental states.[20] At some crucial point of increasing complexity of interaction with the environment, O will need to select one course of action out of many possible depending on the predicted outcome of that action. O will thus need to represent itself as part of the world and represent counterfactual situations (what would happen if it took action A). Such representational capacity seems to be sufficient for O to implicitly distinguish between the subjective and objective order of things - the feature by which Strawson characterizes the transcendental self-consciousness (see p. 4.5).

Another mental capacity that requires O to distinguish between the objective and subjective order of things is the ability to recognize perceptual errors. O can realize it made a perceptual error only if it can conceptualize the perceptual classification *qua* representation. It would be difficult, but perhaps not impossible, to attribute this capacity to a non-linguistic animal that has no explicit means of reference to its representations. The pattern of behavior that O needs to exhibit in order for the ascription of error recognition to be justified is perhaps more complex than the pattern corresponding to planning and counterfactual representation.

As a point of contrast, it is worth emphasising that most bottom-up, naturalist accounts of self-consciousness build on representation of bodily states for the purpose of homeostatic regulation, some primitive sense of self-world boundary etc.[21] What these examples have in common is that they show that the organism needs to represent some properties of itself *as a physical object.* But these examples are relevant only for the empirical self-awareness, i.e. for the capacity to

[19]Note that the rationality constraint on O's action requires intentional access, thus O would satisfy the requirement of elementary self-consciousness as intentional access, described in the previous subsection.

[20]Here I adopt the view of Dennett (1989).

[21]For example Damasio (2012); Dennett (2008).

represent one's objective properties. A simple robot with modular architecture that does not allow for sophisticated integration of information can have representational states directed at itself. Or, to use an everyday example, a laptop turns into the hibernation mode whenever the battery is low, saving the current workspace to a temporary file. The laptop thus represents its internal state (remaining power, CPU heat, etc.), but it is not aware of its representations *qua* representations, which is the mark of transcendental self-consciousness. It may well be the case that monitoring one's bodily states is a precursor to monitoring one's representational states, but they are different capacities.

5.4.4 Self-evidence, cognitive access and redundancy

Assume now that O has the capacity to make a higher order representation of a lower-order conscious state, based on the evolutionary reasons sketched above. What can be said about the relation between the higher-order representation and the object representation? Traditionally, the kind of self-awareness yielded by metarepresentation was thought to have the property of self-evidence: if I am conscious that p, I may form a belief about being conscious of p (schematically, if $\longmapsto p$ then it is possible that a mental state $\longmapsto$ 'I think that p' occurs, where $\longmapsto$ signifies consciousness in the sense of intentional directedness). This conception of self-evidence, with the same representation p occuring both in the higher-order and the lower-order state, is problematic at the vehicle level, although it may be a correct characterization at the content level. If the higher-order and the lower-order states are held to be two different conscious (and hence neural) states, then the formulation is implausible from a naturalistic perspective. If, on the other hand, we wanted to hold that the higher-order state contains the lower-order representation as its proper part, then we cannot explain how our higher-order introspective representations can be mistaken. The conclusion is that the higher-order state should be conceived as a conceptualization or redescription of the lower-order state and self-evidence needs to be reformulated in a neuroscientifically more plausible way. Let me elaborate.

First, self-evidence does not entail incorrigibility: if I believe I am conscious of p, I may be wrong about it. Self-evidence states that if I am conscious of p,

I may form a higher-order belief that I am conscious of p, but the converse is not true: it is not the case that if I believe I am conscious of p, I am indeed conscious p. Schwitzgebel (2011) documents many studies indicating that our introspective access to current contents of consciousness is not infallible (see A for a more detailed discussion of Schwitzgebel's studies). Hurley (1998) gives an example of seeing a die for a short time. I have an indeterminate afterimage of the die, but mistakenly believe that I am conscious of a die with specific number of dots on its side. The object mental state is indeterminate in this respect and my higher-order state mistakenly assigns to it a more determinate content. If we took the higher-order mental state to be simply a matter of adopting a propositional attitude to some content (without redescribing it), there could be no mistake (safe for adopting a wrong propositional attitude, e.g. thinking that one is perceiving p when in fact one is imagining p). Thus to account for the corrigibility of self-reflection, we better understand the higher-order state as a conceptualization of the content of the lower order state (or generally representation of the lower-order state in different format).

Another reason to say that the object mental state is redescribed by the higher-order mental state is that if 1) the higher-order state and the lower-order state were different, non-overlapping mental states, and if 2) the higher-order state was conceived as containing the lower-order state (in the sense that 'I think p' contains p), then the higher-order state would be redundant - it would not provide any extra information to the cognitive system. And a deflationary account of self-reflection at the level of vehicles of content does not make ecological sense. Given that representing information by neural activation incurs metabolical costs, there is a strong reason that such redundancy is unlikely to occur by chance and persist.[22]

But if the higher-order state is a conceptual redescription of the lower-order state, what sense are we to make of the self-evidence claim? Here the only answer, I think, is to interpret self-evidence along the lines of the model presented in 6.1.2. The higher-order state re-presents the lower-order content in virtue of describing

[22]For further argument that metarepresentation (a higher-order representation) should be understood as a redescription of the object representation, see section 3.2.3.

it - most of the time - correctly.[23] If the metarepresentation did not capture the lower-order content correctly, it would not enter the loop of sustained activation with the lower-order content and hence would not itself become conscious - except for the occasional cases where a mistaken metarepresentation is selected to the global workspace because it gains strong support from other currently conscious contents. Importantly, however, the metarepresentational redescription cannot be systematically wrong because then it would not describe the kind of lower-order content it does. For what the metarepresentation describes depends on the patterns of lower-order states and their higher-order conceptualization.[24]

The reason why contents of one's consciousness seem to be always available to reflection is that the higher-level representation is poised to enter the global workspace (consciousness) anytime it is relevant. Modules responsible for the

[23]Correctness, in turn, would have to be cashed out in terms of pragmatic or ecological usefulness. We have developed means of re-presenting features of our mental states (the most prominent feature being the content) in order to be able to do something useful. Communicating our ideas to others is the most obvious purpose of this metarepresentational capacity but this obfuscates the issue by inviting the idea that metarepresentation is just a linguistic articulation of the lower-order content. To avoid this, consider judgements of learning (JOL), a standard example of metacognitive capacity. When studying for an exam, students employ JOLs to decide which topics they need to revise. The judgement of how well a topic has been learnt is arguably a representation of a feature of mental state, hence it is metarepresentational in that it is about representation, not the represented thing. If JOLs were largely inaccurate, i.e. if they could not predict how well the subject will perform at a test, they would not be used for this purpose and hence would not represent the quality and stability of one's knowledge. However, they could still sometimes be wrong, for example when the need for accurate prediction is trumped by the need for favourable self-image or leisure time.

Knowledge of a topic, however, is a dispositional mental state, not an occurent conscious mental state; so it is fair to say that this example does not shed much light on how metarepresentation of conscious states could be wrong.

[24]Consider the following realistic, albeit controversial, example from the classic study on misattribution of arousal by Dutton and Aron (1974). In their experiment, male participants walked over either an unstable suspension bridge (experimental condition) or a solid bridge (control condition). The experimental group showed increased heart-rate and sweating, presumably due to the sense of danger elicited by the dangerously looking bridge. Crucially, males in the experimental group showed a much increased rate of asking a participating female researcher on a date, compared to the control group, the interpretation being that they misattributed the arousal to attraction, not to the dangerous situation encountered earlier.

higher-level representation are already wired to represent particular aspects (contents) of one's consciousness, they don't redescribe it from the scratch each time. In Kantian terms, the mind is already equipped with concepts pertaining to inner sense. These concept are learnt over time, of course, but they can be considered fixed when considering things in the short-term. The reason why self-reflection may seem to be unlimited with respect to the conscious content it can re-present is that many of our self-reflective thoughts are articulated in language. The generative character of language confers much enhanced flexibility of thinking and consequently of self-reflection as well.

Clearly, the account above is only as good as the more general explanation of why some representations are conscious and how they are integrated. Here I outlined the mechanism using concepts from the global workspace theory, but it could also be described in terms of the predictive coding theory, *mutatis mutandis*. The important point is recognizing that the fundamental part of explaining self-reflection is having a good account of the integration unity and of the evolutionary reasons that render the capacity of self-reflection adaptive.

5.4.5 Return of the perceptual model of self-reflection?

Finally, let me address a potential objection to the account above, namely that it falls back to the perceptual model of self-reflection against which I warned in the introduction. After all, if perception is often understood in a simplified way as a cascade of neural processes, each transforming the sensory information at the level below, what is the difference to the suggested model of self-reflection which claims that self-reflective thought is a conceptualization (simply a kind of transformation) of the underlying object state?

The objection invites many responses. First, as argued in the introduction, the perceptual model is misleading when we drawn an analogy between what perception is at the personal level with what self-reflection might be at the vehicle level. At the personal level, the subject *qua* a situated agent indeed perceives things independent of her, but it would be a mistake to assume a homuncular subject at the vehicle level observing his object states. Note that in our model,

the higher-level mental states conceptualizing the information in the lower-level states are not (and cannot be) identified with the subject.

What about the implication that the lower level states are fixed and independent of the higher-level states? To think that forming a self-reflective thought does not in any way change the object thought would indeed be naïve. But our model is not committed to such a view. First, forming a thought about the current contents of consciousness obviously changes the content of consciousness by becoming part of it. And in case of self-attribution of a mental state based on unconscious information (unavailable at the personal level, for example thinking that one is angry), the reflective thought may change the object state in virtue of a feedback from the conscious state to the unconscious object state - perhaps the reflection makes us shift our attention and the anger then may subside, or it can trigger negative associations that amplify the visceral signals of the emotion. Simply put, the general idea that self-reflective thoughts be understoood as redescriptions of the lower-level thoughts does not preclude the possibility of a feedback from the former to the latter.

6 Unity of Consciousness in Cognitive Neuroscience

In this chapter I will discuss those contemporary neuroscientific theories of consciousness that are both influential and show a potential to explain the unity to some extent. Specifically, I will show what the unity of consciousness amounts to in those theories and assess how well they account for the phenomena related to the unity, as discussed in previous chapters.

6.1 Global workspace theory

B. Baars's *Cognitive Theory of Consciousness* (1988) was the first attempt to come up with a detailed model of consciousness that would account for many experimental findings made after the cognitive revolution. The model is supposed to explain why some information is conscious and other remains unconscious despite being discriminated by the brain.

The model first describes the mind as essentially a system composed of specialists which process information unconsciously and in parallel. The architecture supporting parallel processing allows for fast responses to changes in the environment but its repertoire of responses is rather inflexible because the system is limited to the existing connections among input and output specialists. Such a system would have formed only those connections between input and output specialists which make use of an ecologically relevant regularity in the environment to drive an adaptive response, e.g. the smell of food and salivation. Thinking of specialists as localized neural populations, the number of such connections is limited physiologically. Following the later development of the theory by Dehaene, Naccache and Changeux (see below), we shall understand the specialists

more specifically as modules. According to Fodor's characterization of modularity,[1] modules are, among other things, domain-specific (involved in processing only certain type of information and consequently solving only a limited class of tasks) and informationally encapsulated (the module can draw only on the information that comes to it as input, which is usually limited to some domain; it cannot make use of other information that is nevertheless available to the system as a whole). A modular system would thus be limited to reflexive or habitual responses which do not take into account the situational context and are executed even if they are not relevant.

To illustrate this, consider the following, deliberately simplistic example. A rat is approaching food when it suddenly sees a realistic picture of a snake. The visual system feeds an image (a representation) of the snake to right amygdala (one of the areas that are implicated in processing of emotions, notably fear) response which immediately generates fear response and the rat flees. Suppose that further visual processing could let the rat recognize that it was only a picture (e.g. by the fact that changing the viewpoint did not change the image according to the laws of perspective). If this late output of visual processing cannot reach amygdala, the rat cannot overcome this automatic response despite being informed that the snake is not real. Even if the rat's cognitive architecture was modular through and through, there *could be* such a connection that would inform the fear response module of the output of the "image detection" module -

[1]See Fodor (1983). For a more recent critical review of the concept of modularity, see Prinz (2006). Interestingly, Prinz argues to the conclusion that few, if any, mental processes are modular in the sense of Fodor's original definition or even in the relaxed sense of massive modularity proposed by Carruthers (2006). Thus modularity may not be a useful conceptualization of the architecture of the mind, after all. I argued in section 3.2 that modular view of the mind is indispensable for the the concept of neural representation and since the theories presented here all emloy the concept of neural representation, the modular view should be defended. A possible way of defending modularity against Prinz's criticism is to articulate a minimal sense of modularity needed for the concept of neural representation as opposed to the richer and therefore more debatable conception of modularity according to which high-level mental processes, such as object recognition, are modular. To argue that there indeed is a middle-ground of weak modularity entailed in the representation view that is between the strong sense of modularity and rejection of neural representation alltogether is beyond the scope of this work.

but this connection would have to exist beforehand. And there would always be some unconnected modules whose exchange of information could prove useful in novel situations.[2]

Baars therefore proposes an architecture that would keep the benefit of fast, parallel processing while enabling the system to respond flexibly to novel situations. The latter is considered to be the main function of consciousness (see section 3.1 for details). The proposed architecture involves, besides individual modules, a global workspace whose function is to broadcast its content to all modules and let all modules compete for access to it. The purpose of getting access to the global workspace is to make the information available to the rest of the cognitive system. To illustrate the function of the global workspace, consider again the rat from our previous example. If the rat had this global workspace, the "image detection" module could win competition over access to the global workspace, have its message broadcast globally, and if the message reached an appropriate module contributing to action control, the rat could eventually suppress the automatic fear response. Importantly, the global workspace does not have executive function: it is not by itself responsible for accepting one information and rejecting other, nor does it broadcast the winning information selectively only to some modules.

According to Baars, which representations gain access to the global workspace depends on their informativeness, defined as reduction of uncertainty. The more unexpected the information (given the context), the more likely it becomes conscious - an idea that is echoed in the predictive coding theory. The reason for this assumption is that unsurprising features in the environment are already covered in the system's expectations and can be reacted upon by learned, habitual (and hence unconscious) processes. Baars is not very clear on how informativeness is assessed or conveyed, and what, in consequence, adjudicates the competition for

[2]Note that an architecture in which every module were connected to every other module (which would theoretically obviate the communication problem) would be very inefficient since modules would be flooded by mostly irrelevant information, given that modules either send their output to all receivers or none. (This assumption is crucial, for without it one would have to describe a mechanism for deciding to which modules a signal is to be sent, and this in turn would need to be modular.)

access to the global workspace. This point is important, for if the assessment of informativeness were thought to be the task for the global workspace, we would have a theory which is homuncular in disguise.[3] And if the task is assigned to specialists, they would no longer be the fast-but-single-purpose specialists, for they would need to be able to assess kinds of information outside their area of expertise.

The problem of how the competition for access to the global workspace is realized can be partly clarified by an analogy with people in a meeting room deciding on which topics should be on the agenda. Each person, from her limited perspective, evaluates how important a topic is, and the most important topic is thus selected. This process can become quite ineffective as we spend more time discussing what to work on rather than actually working. In the global workspace model, such a process would defeat the purpose of parallel processing by adding constant evaluation on each specialist's agenda. Baars thus argues for a hierarchy of workspaces of increasingly global reach, passing on information that is considered relevant by specialists that work on similar tasks.

Baars's model is purely functional: anything with the same architecture would have the necessary means to exhibit conscious behavior. Dehaene and Naccache (2001) try to flash out the workspace model more specifically in terms of neural architecture. They start with the assumption that the mind is essentially modular and that consciousness is a matter of sharing information across modules. In their view, the global workspace is a dynamic pattern of sustained activation of various brain areas which, on the one hand, are mobilized by attention and, on the other hand, keep other areas in the current global worskpace assembly activated. This self-sustained loop of activation is presumably made possible by widespread and long-distance projections that exist among various (mainly cortical) areas. The theory specifically assumes that these connections are realized by a system of 'workspace neurons' whose role is to modulate activation in the target areas and whose presence therefore limits the scope of representations of which we can become conscious .

[3]Baars explicitly designs his model on the basis of the theater metaphor of consciousness. However, since the audience consists of all specialists (not a homuncular subject), it does not run into the problems associated with the Cartesian theater model.

To illustrate this more specifically, consider the following picture. First, there are always some workspace neurons active in a living, conscious brain. The pattern of activation of the workspace neurons corresponds to what Baars (1988) calls 'context', i.e. desires, goals, dispositions, activated action schemas, etc. This context determines what information is relevant by amplifying those cortical activations that correspond to representations of relevant features. (The relevance, in turn, would have to be cashed out in terms of strengths of preexisting connections among areas.) In virtue of reciprocal connections, the activated areas shape the pattern of activation of the worskpace system and thus change the context in a bottom-up way. If the workspace and a candidate area enter into a self-sustained loop of mutual activation, the activation pattern becomes relatively stable and the representation becomes conscious.[4] Importantly, the workspace system may engage in many such loops at the same time.

As I noted earlier, the theory tries to account for many experimental findings related to consciousness. Let me finally describe a few to illustrate how the theory works. For example, the reason why why we are not aware of the state of our livers is that liver-monitoring interoceptive area presumably does not project to the global workspace (it does not send its output to a wide range of modules). The reason why such a projection has not formed is precisely because the allostatic reaction to a change in liver state is rather independent of the context the organism finds itself in and hence can be carried out unconsciously. Next, we are not aware of briefly presented stimuli because establishing a sustained activation in cooperation with other modules takes some time (hundreds of ms). Last, priming effects exist because the priming stimulus can still influence a wide range of specialist (via strong connections formed during habituation) although it does not get into the temporarily stable loop of sustained activation.

[4]This is an observation, rather than an explanation. We know from experiments that awareness of stimuli requires their presentation for at least 100 ms and also that the onset of awareness, to the extent it can be timed at all, comes significantly later than the disposition to react to stimuli unconsciously. But it does not answer the question crucial for the hard-problem proponent, namely what it is about the sustained global activation that it gives rise to (phenomenal) consciousness.

6.1.1 The access unity according to the global workspace model

How is the unity of consciousness accounted for by the global workspace model? In general, the global workspace model aspires to cover the normative and objective aspect of the unity of consciousness which we identified in chapter 3. What corresponds to the content of unified consciousness at a time is the transient pattern of self-sustained activation. By entering this loop, neural representations become integrated and hence conscious. That they are integrated in the stronger sense of making a coherent content of consciousness is explained in terms of their relevance or informativeness. This concept is thus supposed to do a lot of the philosophically interesting work.

The fact that a group of neural activations forms a self-sustaining loop is explained as those representations forming a coherent set that can be a basis for an ecologically relevent action, given a context. In other words, what makes the neural activations (vehicles of conscious contents) support each other (sustain the activation) is that their association is ecologically relevant. Consider an example from playing a racket sport.When we estimate where the ball will land in order to position our body, we use a limited but wide array of specialist modules: we rely primarily on vision (to see how the player hits the ball and observe its trajectory), we implicitly use our own body schema specific for the sport to anticipate where the oponnent is going to place the ball, given his body position, we use sound cues to differentiate shots like slices, topspins and smashes, and we are aware of our body position so that we can plan an efficient movement. In the context of the sport, these representations are important and hence focused on. Thus the content of consciousness complies with the instrumental rationality constraint in virtue of being realized by a network of connections that have formed because association between the connected areas proved to be useful (ecologically relevant) in the past.

Objectively, the unity corresponds to causal relations among activations of the areas recruited in the global workspace. Dynamic core and information integration theory (see 6.3) provide a more detailed view of how the unity of a pattern of causal relations is to be understood. In this respect, dynamic core and integration theory are compatible with the global workspace theory.

Conceptually, the unity of consciousness follows from identification of consciousness with the global workspace and the fact that there is only one global workspace subserving an organism. Stating this identity is trivial, of course, so to interpret it as a substantial claim we need to consider the reasons that explain why there is only one such system and how its function constitutes consciousness. A standard answer to this is that the brain instantiates unified consciousness because its purpose is to effectively control the behavior of *one* agent/organism. Embodiment and self-organizing character of living things is thus considered a pre-requisite of there being a point of view which, with added complexity, becomes more and more conscious.

Last thing to note is that in the elaborated model by Dehaene and Naccache (2001) the unity is a matter of self-sustained *pattern* of activation, not a matter of activation of a particular brain area. This pattern changes over time as new brain areas are mobilized and old, formerly conscious, areas demobilized, and this change is a consequence of changes in the environment (both inner and outer):

> This active workspace state is not completely random, but is heavily constrained and selected by the activation of surrounding processors that encode the behavioral context, goals and rewards of the organism. In the resulting dynamics, transient self-sustained workspace state follow one another in a constant stream, without requiring any external supervision. (Dehaene and Naccache, 2001, p. 15)

The take-home message here is that those things which are most relevant in the current situation are elevated to consciousness and together form its stream. It ought to be emphasized that *relevancy* is again understood in neural terms as the propensity to enter into a self-sustained loop with current context determined by the current pattern of activation of workspace neurons. This might seem to undermine our control over what information we attend to, but the attention mechanism is, unsurprisingly, considered to be an important part of the workspace system. Again, what the attention is directed at depends on the context.[5]

[5]To be clear, no reductionist theory of consciousness holds attention to be completely free of causal influence of the environment. After all, the reduction consists in showing how high-level cognitive processes such as attention result from causal processes in the brain. Nevertheless,

Since the purpose of the global workspace is to enhance flexibility of behavior by integrating relevant information processed in various domain-specific modules, I take the global workspace model to exemplify the idea of the unity of consciousness as integration. However, this kind of integration unity should be distinguished from so-called object unity, that is the unity of features in a representation of an object. For example, seeing one interpretation of a bistable image, such as the vase/two faces image, is a matter of integrating shape and color in a particular combination. This kind of integration is known in neuroscience as binding and the underlying mechanism may be quite different from the one proposed for global workspace model.

6.1.2 The subject unity according to the global workspace model

Let's now move to the subject unity of consciousness. How does the GW theory account for reportability of conscious states and the inuition that every conscious experience is *my* experience? Reportability could be explained as a matter of the language production module being part of the workspace system, having thereby access to outputs of modules whose activations correspond to current contents of consciousness.[6] If prompted, the language module gets recruited in the global workspace and produces verbal output of the content of consciousness since that is the input it receives when it is a part of the GW.

Similarly, the empirical self is to be understood as a complex of representations realized by various self-oriented modules, e.g. those responsible for autobiographical memory, interoceptive information, theory of mind, etc. Empirical self-awareness arises whenever these modules are recruited by the global workspace. Introspection, for example, could be viewed as a result of the intepretation done

---

the distinction between bottom-up and top-down control of attention that is used in cognitive psychology could be explained in terms of the global workspace somehow, for example as a difference in the way a neural representation joined the global workspace - either as a result of a change in the saliency of a stimulus (bottom-up) or a change in the composition of modules recruited in the workspace.

[6] *Mutatis mutandis* for motor control if the repors are non-verbal, e.g. pressing a button in an experiment to report awareness of a stimulus.

by the theory of mind module as it gains access, through global workspace, to information about one's bodily states, verbal reasoning, episodic memory, etc.

The distinction between empirical and transcendental self-awareness could thus be understood in terms of this theory as follows. Empirical self-awareness is a matter of explicit representation of properties of the self via modules subserving introspection. Transcendental self-awareness, on the other hand, refers to the capacity to form a higher-order thought about current contents of consciousness. Self-reference without identification could then be explained in terms of a higher-order representation being sustained by the activity of the systems in the global workspace in virtue of representing their content. To illustrate, suppose there is a module whose job is to articulate (conceptualize) contents (or outputs, to accomodate for their encapsulation) of other modules. This system will monitor activity of other modules and formulate its best guess of what else is going on in the brain. If the representation expresses the contents of the global workspace (i.e. outputs of the systems of which it is composed at a time), the global workspace will support it - in the same way as neural activity corresponding to a seen object supports the activity of its semantic representation, and vice versa in the case of visual imagery. As a consequence, the higher-order representation will enter the loop of sustained activation and will, in turn, help to sustain that particular assembly of representations. So the output of this monitoring module will become conscious only if it fits current contents of the global workspace, i.e. only if it yields a correct metarepresentation of conscious contents. If the higher-order representation does not reflect the content of the global worskpace, it does not get the support needed for it to become conscious.[7]

[7]As I cautioned in section 3.2.3, the way metarepresentation ought to be conceived of at the neural level renders the distinction between right and wrong higher-order representation of a lower-level content rather metaphorical and possibly misleading. But then, how should we understand the vague expression that the higher-order representation *reflects* the lower-level contents of the GW? I suggest we think of the higher-order representation on the model of abstraction from particular details and extraction of what is invariant across multiple instances of the same content. It is then important to emphasize that the higher-order representation thus conceived can be inaccurate in the same way as, for example, a semantic representation of a seen object may be an inaccurate higher-level representation of the lower-level visual information. What might make the higher-order reflective thought more robust is the fact that were it

A process like this could yield the immediate kind of self-knowledge characteristic of transcendental self-awareness because matching the representations and their metarepresentation is an unconscious process (we are not conscious of the content of the monitoring module until the match is successful, as it were). The contrast is with two representations that can be conscious (recruited in the global workspace) even if they don't match, for example in the case of looking at someone (visual representation) and trying to tell whether it is a specific person we last saw ten years ago (a representation origination in the long-term memory). Furthermore, it explains why the metarepresentation cannot be systematically wrong and hence why we have a strong intuition about the infallibility of introspection.

In the account above, we made a few strong assumptions for the sake of illustrating the basic idea of how a monitoring module and its interaction with the global workspace could yield the phenomena related to the subject unity of consciousness. Perhaps we can now relax some of these assumptions to get a more realistic account. First, it is not necessary that for the higher-order representation to enter the global workspace, it must express contents of *all* modules of which the GW is currently composed. To win the competition for access to the GW, it suffices that the current recruited modules consider the higher-order representation relevant, or, in less metaphorical terms, that the higher-order representation gets large enough neural gain from the transiently stable pattern of activation that corresponds to the global workspace. The neural gain may be large enough even if it expresses the content of only some of the modules currently recruited in the GW. At the phenomenological level, this feature would correspond to the distinction between the center and the periphery of consciousness (center being that the content amplified by a self-monitoring module recruited in the GW). At the cognitive level, the feature could explain the role of metacognition and in-

inaccurate, it would not support the pattern of activation corresponding to the current GW and that, in consequence, the GW would either get reorganized or the reflective thought would dissipate from consciousness. This, of course, is just a speculation based on the tenets of the GW theory.

ner dialogue in shaping the contents of consciousness (directing attention) during difficult problem solving.[8]

Another assumption that might be relaxed is that the process of generating a higher-order representation, which is then poised to become a part of the GW, is modular in the sense that it employs one coding scheme ("language") for all possible metarepresented contents. The motivation for holding the assumption is twofold. First, it conforms to the intuition that metarepresentation must be something like articulating the content in language, as we do in the case of explicit self-reflective thought. Such articulated thoughts are prime examples of metarepresentation. Second, the common coding scheme is what enables integration and redescription (simplification, chunking) of contents of different modalities. For each metarepresentational coding scheme, the smaller its span across sensory or representational modalities, the less likely it is that such a metarepresentation would get a strong enough neural gain to enter the GW. To put it differently, if the metarepresentational format allowed for metarepresenting only some class of contents, the metarepresentation could be recruited to the GW only if the current GW consisted of modules to whose contents the metarepresentation is sensitive.

On the other hand, the multiplicity of metarepresentational formats could be advantageous despite the implied limited span. For the multiplicity would render the awareness of current conscious states less vulnerable to impairment due to neural damage. That is, there could be deficits in awareness of some mental states without any influence on the awareness of other mental states. Many neurological cases, most notably the various kinds of aphasia, can be *prima facie* interpreted as selectiveimpairments of specific kinds of awareness. The idea of multiple metarepresentational formats fits naturally with the predictive coding theory (see the next section) which argues that the brain is organized as a hierarchy of areas in such a way that higher-level areas try to predict the activation (representation) at lower-level areas. However, this predictive coding

[8]See, for example, Diaz and Berk (1992) for the role and development of inner dialogue in problem solving. This line of research is important for the question of how language drives self-awareness as it explicitly draws on Vygotsky's ideas about the significance of langauge and culture in the development of human cognition.

principle is held to be characteristic of *all* brain processes and representations, and therefore does not itself explain awareness of current *conscious* states.

6.1.3 Objections to the subject-unity account

The account of self-awareness presented above is liable to several objections which should be addressed. 1) It seems that the empirical and transcendental self-awareness are conflated or unclear at best. 2) Assuming that the subject unity depends on some monitoring module invites the possibility of dissociation of consciousness and self-consciousness that should be observed if the module is compromised. 3) Third, the ability of the monitoring module to tap on the outputs of other modules seems to imply some common language of thought thanks to which the communication is possible. 4) The functional role of the monitoring module makes it poised for being a bottleneck of consciousness and that is not a desirable feature - in fact, the motivation behind the global workspace theory was to explain how we can be flexible and yet fast in our thinking without having any central executive module.

The conflation between empirical and transcendental self-awareness is only superficial. Empirical self-awareness is a matter of self-related content - ascribing some property to a (fictional) object, the self. Transcendental self-awareness, on the other hand, is in this model a matter of causal structure among conscious contents and their metarepresentation. The *outcome* of this causal structure, namely that the metarepresentation is recruited in the global workspace and therefore conscious, is a case of empirical self-awareness. The possibility of this happening in the way described above (and hence the possibility of the contents of the GW to be metarepresented) is what transcendental self-awareness refers to. Note that the difference is in the architecture, not merely in the aspect we focus on.[9] Empirical self-awareness can be explained in terms of standard feed-forward processing - no metarepresentations nor any matching by loops of

[9]One could argue that empirical and transcendental self-consciousness are the same process, the only difference being that the former refers to the resulting content and the latter to the vehicles of that content. However, empirical self-awareness includes not only cases of reflection on one's conscious contents but also of processes that result in an attribution of a state to the self without that state being conscious (e.g. self-attributing a disposition or intention). Hence

mutually sustained activity are necessary to account for representing a feature ascribed to the self. For example, feeling angry can be a result of simple feed-foward processing of interoceptive information, in a similar way as seeing a banana can be understood as a result of feed-foward processing of visual information.[10] In contrast, transcendental self-awareness with its crucial feature of immunity to error through misidentification could not be realized by such an architecture: the loop of sustained activation between conscious representations and a matching metarepresentation is needed.

The second problem concerns dissociability of consciousness and self-consciousness if the latter is realized by a monitoring module. Modules are usually thought to be localized and therefore vulnerable to selective neurological impairment. According to the model sketched above, such an impairment should lead to consciousness without self-awareness, demonstrated by the subject's inability to report on current contents of consciousness. Some might find such a situation inconceivable, but perhaps it is not impossible. Note that the person could still report on her emotional states and or even her intentions and attitudes in the (self-) interpretive manner typical of confabulation and self-directed application of the theory of mind. For example, self-attribution of positive attitude to vegetarians is not dependent on metarepresentation of my current thoughts, it could well be a folk-psychological inference based on the recollection of my past behavior and related beliefs. The neurologically impaired person could therefore still display a substantial degree of empirical self-awareness. When asked explicitly about the contents of current conscious thoughts, the person would initially fail to report anything and later, after adaptation to the impairment, she could start confabulating some contents.[11] To illustrate what it could be like to be conscious while having the monitoring module compromised, consider the expe-

---

the difference between empirical and transcendental self-consciousness at the level of vehicles of conscious contents *is* more than just the difference between a process and its result.

[10] Although neither case is, in fact, realized by purely feed-forward information processing in the brain, the point is that it could be (as connectionist models suggest), and that the transcendental self-awareness could *not* be realized this way.

[11] An interesting experiment in this respect was done by Flavell et al. (2000) (see section A.3 for a short description). It suggests that children, until some age, do not have the capacity to spontaneously monitor their thoughts.

rience of mind-wandering. Schooler (2002) argues that mind-wandering is a good example of momentary dissociation of explicit self-awareness and consciousness. In mind-wandering, we catch ourselves continuing to do a task while thinking about something else, e.g. reading an article while thinking about vacation. We are conscious of the content of day-dreaming but not of the fact that we are day-dreaming about the content - for if we were, we would presumably stop and return to the task.

Even if the preceding account does not convince the reader that the dissociation is conceivable, the problem could be solved by an account of transcendental self-awareness that does not rely on a modular process. Explanation in terms of a non-modular process also has the benefit of obviating the language of thought problem to which we turn now.

The global workspace model requires that the output of any module that can be a part of the GW can be communicated to other modules. This is one of the reasons that lead Fodor (1975, 2008) to hypothesize a language of thought (LOT) - a medium of mental representation that is propositional and exhibits compositionality (though Fodor's aim is to provide a framework for explaining cognition in general, not consciousness). The LOT hypothesis may seem plausible at the computational level. For example, implementations of the global workspace architecture in software agents by Franklin and Graesser (1999); Franklin (2003) indeed use a common representation scheme to which any specialist piece of code can contribute. If the brain really instantiates the LOT with the decompositional structure Fodor proposes, cracking its neural code would be a major step toward solving the mind-body problem. At the level of neural representation, however, the LOT hypothesis is a lot less plausible. Brain-imaging experiments and connectionist simulations of cognitive processes have shown that representation in the brain is sparse, distributed over a large part of the network, often duplicated and very noisy (see section 3.2 for more details). These features suggest that the neural code is unlikely to have even remotely propositional form.

Could the global workspace fulfill its function even if there is no LOT? I think so, although the price for that is lower flexibility.[12] As Dehaene and Naccache

[12]Cf. (Shanahan, 2005, p. 60) who, in his account of the consequences of the global workspace, argues that "One requirement that any instantiation of a global workspace architecture must

(2001) propose, the message selected for global broadcasting is represented in the format specific to the module of its origin. In that case, the 'listening' modules could pick up only that aspect of the message which they are ready to understand based on the specific connections between them and the signalling module. Thus conceived, modules are still encapsulated and their communication is limited by what kind of information they have learnt to share between one another. But doesn't this thwart the purpose of the global workspace? After all, the message could then be passed to all other modules to which the signalling one is connected even without going through the global workspace. What then is the function of the GW if not integrating information in a common coding scheme? The hypothesized role of the GW that remains functionally important is that one message has the privilege of being listened to by all other modules. This ensures, so to speak, that other information which the modules receive thanks to unconscious, parallel processing will have lower influence on the modules' responses. This does limit the flexibility of the system in the sense that we can be conscious of only those associations of features that are represented by connected modules which are also part of the GW. If, for example, the module for number representation and the module for color representation are not connected, we cannot experience a number as having a specific color, safe in the very abstract sense of entertaining that propositional thought. Arguably, people who do experience such associations (which is an example of synesthesia) have a connection between the two modules.

Besides, to achieve the flexibility needed to solve problems requiring associations that have not been formed as neural connections, the brain can use the representation in natural language. The idea that the flexibility that is distinctive

---

meet is that the information processed by the set of parallel specialists must be coded in such a way as to be 'generally intelligible'. That is to say, an item of information is only worth broadcasting if it has the potential to influence usefully the activity of any of the processes that receive it, and this demands a coding scheme that can be understood by all such processes." I disagree with the last part about a common coding scheme. The broadcast information may influence the activity of the listening modules in virtue of coding schemes specific to each pair of the message's source module and a listening module. This way, each module "understands" only that part of the message that it is disposed to thanks to pre-existing associations.

of human consciousness is bootstrapped by the conceptual structure embedded in natural language has a large support among scholars working on consciousness.[13] To put it briefly, language provides mental scaffolding useful for structuring our thoughts, chunking complex ideas into pieces managable in the working memory, store and represent an outcome of reasoning, and perhaps most importantly: it allows for cumulative development of culture and its transmission over generations.

The last objection to discuss is that the monitoring module would be some kind of information bottleneck due to its capacity to represent contents of other modules. First thing to note is that the monitoring module does not have a *functionally* central role. Conscious processing, defined in the GW theory as communication among modules via the global workspace, can still occur even without the existence of some monitoring module, let alone its actual metarepresentational activity. Granted, absence of the module would preclude the immediate awareness of current conscious states, but there is no reason to suppose that consciousness *simpliciter* requires the capacity to reflect on one's current contents of consciousness. This matter is closely related to the question whether animals are conscious. Denying animal consciousness may be based precisely on the ground that they lack metarepresentational awareness, manifested, for example, by their hypothesized inability to differentiate between what they know and what other members of the same species know. But if we take consciousness to be a natural phenomenon of various depths and grades, it is natural to interpret the capacity for self-reflection as an extension (albeit a very important one) of consciousness as such.

Finally, even if we took self-consciousness to be an essential mark of consciousness as such, we could relax the assumption that self-consciousness is realized by a specific module. To keep things simple, I have so far accounted for the immediate awareness of the contents of ones conscious mental states in terms of a single monitoring module that joins the coalition of GW modules in virtue of re-presenting

[13]Especially theorists who try to put forward a theory of mind that is not representational appeal to the importance of natural language providing the mind with the power of generative syntax. Variation on this idea can be found, for example, in Carruthers (1998, 2006); Dennett (1991); Clark (2001); Lupyan and Clark (2015); Dennett (2008); Sellars (1956).

their content. But what if the sense of transparency pertaining to reflection on one's conscious contents, as well as its constant possibility (Kant's 'it must be possible...') were a result of different metarepresentational processes, each with the capacity to join the coalition of GW modules in virtue of re-presenting contents of some (if not all) of the already recruited modules? In other words, what makes awareness of one's conscious contents possible might not be a specific module, but rather the general process of metarepresentation and mutually sustained activation. (However, this brings forward the difficulties related to the concept of neural metarepresentation described in section 3.2.3.) The next section on predictive coding will further elucidate the idea of self-awareness as a result of a matching metarepresentation.

6.2 Predictive coding

Predictive coding is an increasingly popular framework for understanding the nature of information processing in the brain. Its popularity stems from its general application to all levels of cognition as well as its seamless connection to basic principles of other sciences from information theory and cybernetics to biology.

6.2.1 Main principles

The main idea of predictive coding is that the brain is an inference machine which constantly tries to predict its sensory inputs using models of the world, and updates these models according to Bayes rule. More specifically, predictive coding is a theory which 1) understands all cognition as a matter of Bayesian inference, 2) argues that the brain carries out Bayesian inference by constant feedback between top-down predictions and bottom-up signalling of the prediction error, and 3) outlines a general hierarchical architecture at the neural level that could support the feedback loops and hence Bayesian inference.

To use David Marr's distinction of three levels of analysis of an information processing systems, the idea of perception and cognition as Bayesian inference is a description at the computational level, and the predictive coding theory as such is a theory describing the algorithmic and implementation level. As Friston (2010)

and Clark (2013) note, the idea that the brain is an inference machine goes back to Helmholtz, and the predictive coding theory proposes a testable hypothesis of an architecture that could realize it. Let me elaborate on these points.[14]

The rationale for Bayesian inference is that by perception we obtain only partial information (e.g. we see only the side of objects facing us). The same information (here: perturbations of receptors in sensory organs) may have originated from different external sources. This implies pervasive uncertainty that can nevertheless be resolved by probabilistic reasoning: assume that interpretion which is best supported by the information at your disposal given your prior knowledge of the world (i.e. what is the prior probability distribution of the possible causes of that information). The optimal way to resolve the uncertainty is thus combining the novel information with previous knowledge of possible sources of that information. In terms of Bayesian inference, the posterior probability is the product of the likelihood function (function of conditional probability of a hypothesized source given the observed data) and prior probability distribution over possible sources.

To illustrate Bayesian inference, consider an example from Ma et al. (2013) about collecting your suitcase at a baggage claim at an airport. Assume that 1) the first suitcase appearing on the conveyor belt looks like yours, 2) you know that you have a common black model that looks similar to 5% of all suitcases, and 3) you know there were 100 passengers in the plane (thus assuming 100 pieces of luggage). You want to estimate the probability that the suitcase is indeed yours. The solution is to combine the probability distribution of the suitcase being yours *prior* to your observation (i.e. 1/100 chance it is yours, 99/100 it is not) with the likelihood function that specifies the probability of your observation (the first suitcase looks like mine) for each relevant state of the world (which is 1 if the suitcase is in fact mine and 0.05 if it is not mine). According to Bayes rule, the posterior probability of a hypothesis given some observation is expressed as:

$$p(\text{mine} \mid \text{similar}) = \frac{p(\text{similar} \mid \text{mine})p(\text{mine})}{p(\text{similar} \mid \text{mine})p(\text{mine}) + p(\text{similar} \mid \text{not mine})p(\text{not mine})}$$

[14]There are many more or less detailed review of the predictive coding theory, for example Clark (2013); Friston (2010); Penny (2012). I encourage the reader to consult these for further details.

As more suitcases come out, the probability increases as the prior probability distibution over the two states of the world that interest me (i.e. mine vs not mine) increasingly favors p(mine). This captures the idea of updating ones beliefs according to new evidence which in turn leads to a different probability estimate given the same evidence. In our example, having found out that none of the first 50 suitcases were mine, my estimate that the next similarly looking suitcase is mine will be higher than in the first case. The basic tenet of predictive coding is thus that perception as a process resulting in knowledge of the state of the world is achieved through a process following this Bayesian logic: the brain combines the sensory input with its model of the world (knowledge of which states of the world could cause that input) to yield a representation of what state the world is in.

To emphasize the problem of partial information, consider the uncertainty or noise inherent in visual perception. The observation, expressed in our example as 'looks similar', depends on our visual acuity, occlusion by other objects etc. Given that 5% of all suitcases are of the same model as mine, I may consider other models to look similar if those models exhibit the same features as my model does *in the set of features registered by me in the current conditions.* That is, from large distance I might be able to distinguish only the color, thus increasing the ratio of similarly looking (in terms of color only) suitcases to, say, 10%. Thus our perception should be sensitive to the uncertainty in the underlying data as well as to previous knowledge. This motivates Bayesian inference as the model of perceptual processing.

The baggage claim example involves conscious inference. The predictive coding theory holds that all our cognition follows this Bayesian rule, including the most basic feature detection taking place at the lowest levels of the sensory processing hierarchy. The general idea is that the brain tries to predict the neural activity generated by sensory receptors by a hierarchy of predictive models.

Moving from the computational level to algorithmic and implementation levels, the most illustrative way to motivate the predictive coding theory is to consider the trivial fact that the brain does not have a direct access to the world. The brain is in touch with the world only through senses that transduce physical

energy into action potentials. Thus it must somehow make sense of the world from patterns of neural activation originating in the sensory organs. It is argued that under the constraints that organisms face, the best way to achieve this is via a hierarchy of information processing areas in which one area tries to predict the neural activity of an area lower in the hierarchy. That is, neural connections in the higher area encode a predictive model of the activity of the lower area. The prediction from the higher area is then matched against the actual neural activation of the lower area and an error signal is fed forward to the higher area, leading to adjustment in the model so that it better fits the actual data. In terms of Baysian inference, the generative model of the higher area is the prior, the actual neural activation of the lower area is the observation, and the prediction error serves to update the higher-level model, i.e. to compute the posterior. This basic loop is iterated across all levels of sensory processing, so that each area predicts the activity of a lower area and sends prediction error to a higher area. The ultimate aim of information processing is then minimization of the overall prediction error (summed across all areas). Minimizing the prediction error is equivalent, given the proposed architecture, to finding the most probable cause of the sensory state and this, in turn, corresponds to knowledge of the current state of the world. More precisely, epistemic states are generated prediction error minimization in the short-term, while long-term error minimization ensures that the organism takes life-preserving actions. This specification comes from the formulation of the predictive coding theory that builds on the so-called free energy principle.[15]

The free energy principle has a number of equivalent formulations. For our purposes, it suffices to mention two. First, the principle state that any self-organizing, biological system will maximize the extent to which sensory data (evidence) conform to its model of the world, i.e. it will maximize the fit of the model. This can be done in two ways: by updating the model so that it conforms to the sensory evidence, or by taking actions which in turn (after a feedback from the environment) will produce sensory evidence that is in line with the selected model. The former encompasses the meaning of perception while the latter shows

[15]See Friston (2010).

how action can be cashed out in terms of Bayesian inference following the same principles as perception. The free energy principle is thus a natural fit with the embodied mind theory.

Another formulation of the free energy principle relates predictive coding to homeostasis: any self-organizing system seeks a long-term series of states with low overall entropy. That is, the system will "minimize the long-run average surprise of sensory states, since surprising sensory states are likely to reflect conditions incompatible with continued existence."[16] Avoiding surprising states means that the system will often be in a small number of possible states and only occasionally in a large number of states (the set of possible states that an organism can be in is defined by its phenotype). Again, the system will avoid surprise if it has an accurate model of the world, where a model of the world means, more specifically, a model of the dependencies among actions, hidden states of the world, and sensory states.

The free energy principle is sometimes criticized building on an idea that the best way to minimize the prediction error would be to shut off our senses as much as possible, e.g. by staying in a dark room, so that the brain can predict absence of sensory perturbances (e.g. silence and darkness) with near perfection.[17] This challenge is usually answered by arguing that we have strong (possibly innate) priors (expectations) that render this isolationist strategy undesirable (defined as leading to greater long-term average of surprisal). For example, Clark (2013) argues against the dark room problem by emphasizing that we have learnt to expect the environment to be changing and to seek unevenly spread resources to sustain our lives. Staying in a dark room goes against this fidgety prior.[18] Further discussion of this challenge to the predictive coding theory is beyond the scope of this thesis. However, it is important to keep in mind that the theory explicates volitional states as the need to obtain information conforming to our priors, or

[16]Seth (2015) Surprise, or surprisal, is an information-theoretical term for negative log probability of a state, i.e. the less likely the occurence of a state, the more surprising.

[17]Cf. Clark (2013); Seth (2015)

[18]Note that this explanation works only insofar action is understood as ultimately motivated by obtaining evidence conforming to our prior model. A completely passive system would eventually habituate to the dark room environment and updated its fidgety model to a more still version.

as the need to attenuate prediction error pertaining to our interoceptive states. For example, the disposition to feel hungry and its acute phenomenology would probably be explained in terms of a large prediction error coming from the low-level interoceptive area regardless of the actual top-level prediction of the state itself. (To speculate further, the reason why harmful states of the body drive appropriate actions rather than habituation to the pain could be that harmful states generate essentially random (and hence unpredictable) neural activity.)

The standard informal interpretation of predictive processing is that the higher level model represents causes of the lower level activation.[19] The higher an area is in the hierarchy, the more abstract features (the more general causes) it represents. The importance of Bayes rule in updating the models is that it allows the brain to infer the causal structure of the world without any prior knowledge of it.[20]

From the perspective of machine learning, updating internal models via Bayes rule would be one approach to achieve unsupervised learning. Initially, the model would predict just random noise but iterated Bayes update across a variety of inputs would eventually lead to recognition of the statistical regularities within the data and consequently to meaningful classification. Given a fixed input data sample, the complexity of inferred causal structure depends on the number and the size of levels in the predictive hierarchy.

[19]Technically, the higher level encodes statistical regularities found at the lower level. However, it is not inappropriate to understand these encoded regularities as the system's representation of causes (cf. Hume's account of causal relations). This is perhaps more intuitive when considering the higher, semantic level of information processing, rather than the lower level where the 'causes' (regularities) are features of percepts of which we are not conscious as such, e.g. edges, color contrasts or syntactical properties of an utterance. In a sense, the concept of apple, for example, is a generative model predicting co-ocurrence of many physical properties, such as spherical shape, stalk, limited range of colours, sweetness, etc., as well as various affordances (edibility) and sensorimotor contingencies. These properties in turn are models of co-occurrence of some lower-order properties.

[20]This is not to imply that the newborn brain is a complete *tabula rasa*, unconstrained by any implicit knowledge or preferences for certain types of stimuli (cf. Karmiloff-Smith (1992)). Indeed, it is likely that the infant brain harbors some innate models which facilitate learning in the early stages of human development. However, predictive coding implies that these innate models should be changeable if the environment does not support them.

The last thing to consider, before we turn to the unity of consciousness under predictive coding, is the principle of precision weighting. To repeat, cognition is a result of combining two sources of information - the prior generative model and the actual signal. Which of the two will have relatively higher influence on the posterior estimate depends on relative precision of the two sources of information. Precision is defined simply as inverse variance of the prediction or the prediction error.[21] Consider a slightly adjusted example from Penny (2012) about estimating where a tennis ball will land while receiving a serve. I have prior beliefs about where my opponent is likely to place his serve, e.g. based on his previous serves. I also have visual information about the trajectory of the ball from which I can generate an estimate of the landing point. Now, the faster the serve, the less precise my perception of the trajectory, hence the less precise the estimate of the landing point based solely on this source of information. In consequence, my final estimate of where to move to return the serve will be more influenced (biased) by the prior probability distribution learnt during the course of the match. On the other hand, if the serve is very slow, my visual perception of the trajectory is acute (less noisy, more precise) and I would, unconsciously, take the prior beliefs out of consideration and rely only on the visual cue. As the range of velocity of the ball is continuous, so is the precision of the observation estimate. Likewise, the more predictable the play of my opponent, the less the decision where to position myself is guided by the visual cue (until my opponent notices it and tricks me by playing the opposite of what I expected).

6.2.2 Attention as precision-weighting and related problems

Let's now move from perception to consciousness. It should be emphasized that predictive coding has been proposed as a model of cognition, not as a theory of consciousness. Nevertheless, as the predictive coding theory gained on popularity, some theorists tried to extend the theory to account for consciousness as well. Hohwy (2012) speculates that the content of consciousness corresponds to the best fitted model:

[21] See Penny (2012) for corresponding mathematical formulation.

> The core idea is that conscious perception correlates with activity, spanning multiple levels of the cortical hierarchy, which best suppresses precise prediction error: what gets selected for conscious perception is the hypothesis or model that, given the widest context, is currently most closely guided by the current (precise) prediction errors. (Hohwy, 2012, p. 5)

In most of the cited article, Hohwy focuses on attention rather than consciousness. Attention is articulated as context-dependent precision weighting. The underlying motivation is that besides the first-order problem of perceptual inference, which is that of prediction error minimization, the system must be able to estimate precision of its own models and adjust it according to the context since the reliability of feature detectors varies with external as well as internal conditions (for example, fog or dizziness modify (decrease) the precision of visual perception). To take an example related to attention, consider the famous invisible gorilla experiment.[22] If the task is to count how many time a basketball is passed among a group of players, the precision weight of the prediction model for object motion should increase, relative to its baseline and to other visual models. This ensures that the models higher in the hierarchy (e.g. for counting) will be influenced more by motion detection and less by potentially distracting stimuli, such as the walking gorilla. As the example suggests, precision weighting is a way to represent relevance of features in the environment (including one's own cognitive states) and the optimal precision weighting depends on the context (including one's goals).

One thing that is not entirely clear in this account of attention is whether precision weighting is the same thing as the older concept of attentional amplification. On the one hand, Hohwy's account focuses on the role of precision weighting in marking the salience or relevance of a feature: by increasing the precision of a prediction error, it is ensured that the feature responsible for the error will have a greater influence in the cascade of information processing. This is the

[22]In the experiment done by Simons and Chabris (1999), people were asked to count the number of basketball passes made by a group of people on a video. In the middle of the video, a man dressed in gorilla costume walks through the group and thumps his chest while facing the camera. About half of participants miss this event entirely.

role of attentional amplification as well. On the other hand, precision weighting encodes the reliability of feature detectors given a context. Intuitively, I can attend to a very unreliable channel, for example to vision when it is foggy. Granted, the currently unreliable channel is perhaps more reliable if attention is directed at it; but it would still be important for the system to track its low reliability in the current context. Given that consciousness corresponds to the hypothesis most informed by precise prediction errors, either vision must be assigned high precision despite its low reliability (it is foggy) or its precision weight is increased only a bit (relative to a case when it is foggy but we do not attend to visual information) - but then there are likely to be more precise sources of information that should consequently be elevated to consciousness (e.g. sounds, touch, etc.) by attentional capture. To put it differently, precision weighting as a marker of relevance seems to be incompatible with the other purpose of precision weighting, namely keeping track of the reliability of the weighted feature detectors.

Ransom et al. (2017) voice similar concern about the purported explanation of attention simply as precision weighting. There is a particular aspect of attention, they claim, that cannot be explained in the way Hohwy (2012) and others propose, namely selective, top-down controlled attention. A somewhat detailed discussion of the issue is useful since the argument generalizes to possible accounts of mental action in terms of predictive coding. Consider a visual scene with two overlapping images. Such visual experience occurs, for example, when we are looking out of a window with a light source behind us: we can see both our reflection in the window and what is outside, and can attend only to one of the two at a time. That we cannot attend to both is neatly explained by the predictive coding theory as a result of not having priors for such overlapping images, hence the prediction error pertaining to the visual scene is always maximally explained by one interpretation. Top-down change in attention is explained as "the representation of an expectation concerning the *precision* of the prediction-error signals".[23] But in the case of overlapping images, precision of the signal corresponding to each interpretation is the same. Certainly, the system can *assign* more precision to features of one interpretation and thereby make it conscious

[23]Ransom et al. (2017)

(a predictive-coding way of expressing the idea of attentional amplification), but then shifting attention cannot be interpreted as a change in expectations about precision, since the precision of both images (interpretations) remains the same.

To save the attention-as-precision-weighting interpretation, the authors consider the suggestion that as expectations about precision lead to signal amplification, the amplified prediction error indeed becomes more precise (less variant). This creates a positive feedback loop, leading to self-fulfilling prophecy of the form: what I am going to attend to will be worth attending to (because focusing my attention on it will make the signal more salient). But then, switching to the other interpretation (changing attention) cannot be interpreted as expecting higher precision of the now unattendend interpretation, since that positive feedback loop increased relative precision of the current interpretation. In other words, the more I attend to something, the less likely I should be to switch attention, other things being equal.

Setting these technicalities aside, the general problem is how to explain mental action. In the predictive coding theory, action is understood as active inference, i.e. reaching a match between prediction and the prediction error not by updating the predictive model but rather by bringing about changes in the world so that it conforms to the model which is currently selected. Standing up, for example, is a matter of predicting the proprioceptive feedback pertinent to standing upright and suppressing, momentarily, the proprioceptive prediction error via precision weighting (otherwise we would not stand up, only realize that we still sit despite our intention to stand up). Mental action, however, is more complicated in that the changes in prediction errors do not come from the agent effectuating changes in the world, but only from changes in the focus of attention which are nothing but changes in precision weighting. As a consequence, top-down driven changes in expectations of precision would be self-fulfilling and therefore completely stable. Thus any change in attention would have to be caused by an externally driven change in expectation of precision - which is to say that no attentional shift should properly be regarded as top-down. But then it seems difficult to account for experiences in which we can, at will, change the focus, for example when we switch from one interpretation of a bistable image to another.

6.2.3 Bodily self-awareness

With its unified account of action and perception that is congenial to the embodied cognition framework, the predictive coding theory provides a convincing account of bodily self-awareness. The core mechanism behind the bodily self-awareness is a strong coupling between (physical) actions and corresponding interoceptive and proprioceptive signals. Self-generated movements are initiated together with predictions of proprioceptive signals to ensue so that we can make quick adjustments and corrections to the movement while the limbs are still in motion. A phenomenologically telling example is that of walking up stairs, absentmindedly, and trying to take one more expected step that is nevertheless not there. Even before the foot reaches the floor, we become alerted thanks to the mismatch between the (unconsciously) expected proprioceptive feedback from the foot touching a stair and the actual feedback.[24] This idea has been around for a long time (allegedly since Helmholtz) under the name 'efference copy'. It was theorized, and later corroborated by experiments, that efferency copy is the mechanism responsible for a variety of phenomena, for example saccadic suppression of image displacement, attenuated sensitivity to self-induced tactile stimuli (e.g. tickling) or motor adjustments.

The efference copy mechanism is essentially a predictive model of sensory states given a particular action. The predictive coding theory states that this mechanism is ubiquitous rather than specific to only some cases. This invites explaining bodily self-recognition as a result of predicting sensory states based on acting upon one's own body. Touching one's own body produces a multitude of temporally congruent sensations - for example visual perception of two objects in contact (e.g. one's arms and legs), correspoding tactile sensations, and a series of proprioceptive sensations as the movement unfolds. This regular association will lead to recognition of one's body as that object in the world which, when

[24]Note that the expectation becomes phenomenologically salient only when it is not met, which in terms of predictive coding is when the prediction error is large. This illustrates the idea put forward by Hohwy (2012) and others that the content of consciousness correlates with the models that are at a time in charge of explaining large and precise prediction errors.

acted upon, gives a specific type of feedback. In contrast, watching two external objects in contact will not produce any regular tactile sensation that could be reliably mapped onto what one is seeing. The sense of touch and proprioception is not, of course, the only information channel that can support self-recognition. Recognizing oneself in the mirror, for example, could possibly be achieved solely by coupling of visual sensations with motor commands: I am that face in the world whose changes can be predicted entirely thanks to motor commands (while changes in other people's faces can be predicted far less reliably and only mediately via knowledge of context, the theory of mind etc.).[25]

This account has some important implications: 1) there are many subsystems (each represented by specific prediction model) contributing to the seemingly unified bodily self-awareness, 2) self-recognition is a matter of specific structure of information processing (matching motor predictions with sensory feedback), rather than specific kind of information, and 3) our representation of bodily self is probabilistic and hence malleable. Let me elaborate on these claims.

The first two points are closely related and are explicitly mentioned in the accounts of self and self-recognition put forward by Apps and Tsakiris (2014) and Seth (2013). Their intention is to explain, by one encompassing theory, a variety of self-related phenomena known in cognitive psychology, e.g. the sense of agency, the sense of body ownership, or self-recognition. The problem is to explain how come that we ascribe a specific property to a perceived object, namely the property of 'being me'. This is a simpler problem than that of accounting for the unified sense of self because recognizing some object as being me (my body) seems to presuppose that I have a prior sense of self (transcendental self-consciousness, see section 4.5). Under predictive coding, the concept of an enduring bodily self would correspond to some predictive model that is higher in the hierarchy than the contributing modality-specific self-recognizing mechanisms. As (Apps and Tsakiris, 2014, p. 93) put it, "the representation of self must be hierarchically distributed and recruit in all unimodal systems that register the consequences of self-made acts."

[25]See Apps and Tsakiris (2014) for further details.

The common denominator of the various self-recognizing mechanisms that ensures their integration in a more abstract representation could be the temporal congruence of sensory, interoceptive and motor information that is uniquely present in the case of observed interaction with one's own body. That there is a multitude of modality-specific mechanisms for self-recognition is supported by the fact that there are many brain areas implicated in tasks involving some representation of the self.[26] Further support comes from behavioral evidence showing that animals and children can succeed in one self-recognition task while fail at another, thus suggesting that self-recognition is modular. For example, two-year olds may pass the mirror test but often fail to recognize themselves in a video recording, even if played immediately after the recording.[27] This particular example illustrates the importance of synchronous sensory and interoceptive feedback on action, at least in the early stages of developing a more abstract and robust concept of bodily self.

Regarding the third point, saying that the representation of bodily self is probabilistic is trivial insofar as every representation is probabilistic according to the PC theory. Note that the concept of probabilistic representation entails the possibility of error, in this case misidentification of an object as me. Thus we should not expect this to be an account of the problem defined in 5.2, namely how the self-ascription of properties that is immune to misidentification is possible. Still, experimental manipulations of the sense of self are interesting since they show that bodily self-awareness is far from self-evident.

Consider the classic example of the rubber hand illusion, in which a rubber hand is placed on a table in a position similar to the subject's real hand that is occluded from her vision. If both the rubber hand and the real hand are stroked with a paintbrush, the subject is likely to report that she experiences the rubber hand to be hers. Importantly, the stroke must be simultaneous (to produce a tactile sensation in the real hand and a congruent visual perception of the rubber hand) and have the same direction. When asked to point to her hand, the subject is more likely to point to the rubber hand's location. As

[26]See Seth (2013) for a list of the implicated areas.

[27]See Gopnik (2009) and Siegler et al. (2011) for further discussion of the development of children's concept of the bodily self.

Suzuki et al. (2013) demonstrated, the illusion can be induced even without tactile mediation. In their experiment, the subject watched a virtual 'rubber hand' changing color from normal to reddish either in sychrony or asynchrony with the subject's heartbeat. As in the original experiment, the illusion occurred only in the synchronous condition. Furthermore, the researchers found that the strength of the illusion correlated with interoceptive sensitivity, measured as the ability to detect one's heartbeat.

The PC interpretation of the experimental findings is that the synchrony increases the likelihood of the rubber hand being mine to the point that it overrides the prior representation of the location of one's hand that is hidden from one's view. Generally, any perceived object (including sounds and even thoughts) is a candidate for being me (my part) and the posterior probability can be manipulated by tinkering with the variables that the predictive model takes into account. The PC theory also implies that the sense of self may arise from other sources than just the integration of sensori-motor efference and reafference - any self-related information could shape the probability distribution of an object being me.[28]

6.2.4 Transcendental self-consciousness

Does the predictive coding theory offer a convicing account of transcendental self-awareness as well? As mentioned in the previous section, to recognize something as my hand, it seems necessary that I have a sense of myself as the common subject of the recognitional activity and the sensory perception of the arm. In Kant's terms, it presupposes the thoroughgoing identity of the 'I'.

Let's start by repeating a point made earlier about consciousness in the PC theory. Contents of consciousness correspond to that 'hypothesis' that currently best explains precise prediction errors. To put it less technically, we are conscious of the brain's best interpretation of those perturbances in its own activity that are weighted as highly informative. This implies, together with precision being relative and continuous, that consciousness is like a dynamic field, with center and

[28]To illustrate this, Apps and Tsakiris (2014) discuss experiments showing how performance in self-other face recognition task can be influenced by self-related primes or cultural differences.

periphery, of different width and focus depending on the distribution of precision weights.[29] The difference between conscious and unconscious representations is then only a matter of degree. The idea that consciousness can vary in its scope and intensity is perhaps phenomonelogically intuitive if we consider the contrast between concentrating on a simple single domain task (e.g. hitting a baseball) and relishing the wide context of being on vacation.

Coming back to the unity of consciousness, the identity of the 'I' across representations in the PC theory can be understood as the identity of the cognitive system which generates a hypothesis about the current state of the world with the greatest scope (integrates information from most sources) and likelihood. That is, the synthetic unity of apperception would correspond to informational relations among conscious contents at a time to the effect that they are evidence for the winning hypothesis. Each predictive model synthetises, to use a Kantian term, representations encoded in the lower-level areas whose activity it tries to predict and whose activity, in the long run, determines its content by some learning mechanism. The unity of apperception then goes as far as the integration of various predictive models does - how well the models minimize prediction error by what we experience as a coherent and integrated representation of the world.

Now, we may ask why is it that we experience a unitary, coherent representation of the world rather than a set of unrelated contents that individually explain away predictions errors in particular domains. To ask differently, is there a principled reason why the brain should instantiate just one hypothesis that integrates all representations?[30] I think there is. Recall that the free energy principle implies minimization of the sum of prediction errors across the brain. Two competing hypotheses (with different predictions) would generate large prediction errors. That is, unless the brain consisted of two functionally independent

[29]If we assume that precision weights always sum up to a constant, we can conceptualize focused consciousness as cases where only few models are assigned with high precision. On the other hand, 'broad' consciousness corresponds to precision weights being assigned evenly across many models (mind-wandering could be a good example here). The assumption that precision weights always sum up to a constant is equivalent to saying that attention is a finite resource that can be allocated in different ways.

[30]Cf. split-brain patients and the hypothesis of multiple centres of consciousness.

parts,[31] each minimizing, so to speak, its errors by its own hypothesis. The reason why functional independence is highly unlikely is that the brain is a control system for *one* organism. The free energy principle implies that the optimal way for the organism to minimize surprise is to make sure that its actions are coordinated and consistent to the highest degree possible. This renders functional independence suboptimal. In other words, the physical and ecological unity of the embodied agent makes it optimal (and hence selected for during evolution) that the behavioral control system will be integrated so thoroughly that it would be interpreted as a single subject from the intentional stance.

The crucial part of a potential PC account of the unity would thus be the explanation of how various predictive models are recruited to form one hypothesis. The PC explanations are plausible insofar as they concern a phenomenon that can be understood via hierarchical processing (with a tree-like structure branching to lower levels) of prediction errors. Indeed, most illustrations of the PC architecture assume acyclical structure of processing nodes. Nevertheless, since there is no single place in the brain that would function as the final synthetiser (a place in which all processing pathways would converge), the overall winning hypothesis about the state of the world, which presumably corresponds to the content of conscious experience at a time, must be realized by some assembly of prediction models that cannot be hierarchical. How is this assembly formed and what is the nature of the information flow in this assembly that it yields one winning hypothesis despite lacking hiearchical structure terminating in one node? This is the crucial question that the predictive coding theory has not yet fully answered. Note that the global workspace theory faces a similar, if not the same, challenge.

A preliminary answer could be that higher-level prediction units form connections (possibly bidirectional) whose strength is proportional to the extent to

[31] By 'functionally independent' I mean a situation in which the activity in one part has no causal influence on the activity in the other part of the brain, at a time. This might be specific to a situation (imagine a contrived example of doing different task with each hand with relevant information coming in by that very hand), hence *functional* independence. Technically, functional independence would correspond to the situation that there exists a division of the brain to distinct parts A, B such that the Markov blanket for any node in A is part of A, and similarly for B.

which the connection between the hidden causes that these units represent is ecologically important for the organism. For example, a professional poker player might have a relatively stronger connection between her probabilistic reasoning and social cognition for the sake of better betting decisions. Or, to take another example, humans may have weaker connection between social cognition and the sense of smell, relative to dogs, because oflactory cues are rarely relevant for social behavior among humans. (This is essentially the principle of neuroplasticity or neural darwinism that was identified in section 3.2 as the mechanism that can explain why a particular pattern of activation is the vehicle of a particular content.)

An important implication here is that the overall hypothesis is constrained by evolved connections among high-level areas which encode general and more abstract aspects of the world. In Kantian terms, our conception of the world is constrained by the categories (and other, less fundamental concepts) that nevertheless carve nature at its joints.[32] Plausibly, this constraint can be overcome to some extent by language. The ability to explicitly articulate our thoughts can enable us to go through various hypotheses and test them in our minds, without acting on the world according to the hypothesis that would have won in a less flexible cognitive system.[33] Language would thus allow areas that have only weak connections to jointly shape the posterior probability distribution.

Lupyan and Clark (2015) suggest that the mechanism behind this is again precision weighting which effectively marks the relevance of processed features. Note that context-dependent precision weighting is also the core of the account of attention put forward by Hohwy (2012). The idea thus is that language is a tool co-opted for setting and manipulating the context that shapes, via precision weighting, the relative importance (and hence impact) of various sources of in-

[32]Another important implication of the predictive coding theory that is not echoed in Kant's theory is that the set of categories by which we structure the world depends not only on the causal structure of the world and our learning mechanism (Bayesian inference) but also on our phenotype which enables us to pick out statistical regularities only in the dimensions in which we are open to the world. The world is different for creatures that have different senses or whose lives unfold in different temporal and spatial resolution.

[33]Dennett (2008) explains this flexibility-through-language in his account of humans as Gregorian creatures.

formation. It could be argued, however, that this is not much of an explanation, since the heavy and philosophically interesting work is just deferred to another unclear concept that is putatively responsible for the synthetic unity, namely context. The predictive coding explanation would thus be more telling if we had a clear idea how context is instantiated in the brain and how it affects precision weighting at the global level.

6.2.5 Subject unity of consciousness

So far I have outlined an account of the integration unity and bodily self-awareness. Let's now turn to the subject unity, understood as the ability to represent all my conscious contents as mine. As Kant and Shoemaker emphasized, the conscious subject must recognize its identity across mental states without identifying itself via recognized properties - otherwise it would face an infinite regress of justifying the knowledge that the properties used for identification are hers (see 5.2). The efference copy mechanism described above (and its predictive coding generalization) provides an explanation of bodily self-reference without conscious identification. The qualification 'conscious' in the previous sentence is necessary because, at the subpersonal level, the mechanism *does* tag some perceived object as being one's body (hence identifies it) in virtue of recognized properties - namely the temporal congruence of exteroceptive and interoceptive information. However, insofar as the mechanism is unconscious and cognitively impenetrable, the sense of identity of the self is simply given or non-inferential (relative to a conscious inference) at the personal level. Is self-knowledge generated this way immune to misidentification though? Clearly not, as the rubber hand illusion shows. However, identification of a hand, voice, or face as mine is not supposed to be errorless. What needs to be immune to error due to misidentification are statements involving the thinking subject, e.g. 'I think this hand is mine.' (rubber hand illusion) or 'I think I am controlling the motion of this cursor.' (manipulating the sense of agency). So, the subject must know the identity of himself as a thinker and the self-reflecting subject.

Taylor (2012) proposes a model explaining this immediate knowledge of the identity that is congenial to the predictive coding theory although it is not expli-

cated in its terms. He envisions a model of attention control mechanism which not only amplifies outputs of attended channels but also sends an efference copy to 'owner' module which then anticipates the about-to-be-conscious state. If the prediction generated by the 'owner' module matches the result of the attentional shift (i.e. the conscious state that follows), the owner module tags, so to speak, the resulting conscious state as mine. In terms of predictive coding, the CODAM model (COrrolary Discharge of Attention Movement) thus explains the known identity of the self-reflecting subject and the lower-order representing subject as a consequence of matching the predicted neural activity resulting from shifting one's attention to the activity that actually ensues. Taylor claims that the proposed attention control architecture explains Shoemaker's immunity to error through misidentification because I cannot be mistaken, thanks to the attentional efference copy, that it is *me* who turned attention to something and thereby became conscious of it.

Couple of qualifications must follow. First, Taylor's theory would be credible as an account of the pervasive sense of ownership of experience only under the sparse view of consciousness according to which we are conscious only of that what we pay attention to.[34] In that case, attention is implicated in any conscious state and since directing attention always involves sending an efference copy, his theory would account for the pervasive sense of ownership of experience or implict self-awareness. Second, it should be added that the distribution of attentional amplification signal, efference copy of which goes to the 'owner' module, may be an unconscious, bottom-up driven process. So the efference copy must be generated even in cases of surprising stimuli catching our attention - otherwise we would not be sure that the surprising experience is ours, which is obviously false.

Such a conception of self-awareness that is based on working of a certain module or mechanism naturally invites the possibility of malfunction. So if there really is this 'ownership' module that receives the efference copy of attention movement, we should find cases when this mechanism fails. Taylor argues that

[34]See section A for details about the distinction between sparse and abundant view of consciousness.

schizophrenic experience can be interpreted as a consequence of a failure of the attention control system. For example, the delusion of implanted thoughts could be explained as a failure of the ownership module to predict the ensuing content of one's own silent rumination (that is presumably carried out as attentional shift). Importantly, the PC theory offers a similar account of the delusion of hearing voices: it could result from a failure to attenuate (predict) phonological representation of one's own rumination which is then interpreted as an external voice (for the 'internal voice' is usually very predictable and thus does not generate large prediction error).[35]

Taylor articulates his theory in terms of a modular architecture, explicitly referring to the ownership module as instantiating the predictive model of attention movement. The ownership module thus should be able to predict the activity corresponding to any conscious representation. It is very unlikely that any such module exists, however - at least if we take the modular characterization seriously. Even according to the relaxed specification of modularity by Carruthers (2006), and *a fortiori* the original formulation by Fodor (1983), modules are localized and their function is domain-specific, not global. That means that each module processes a limited range of information and yields limited outputs. Consciousness is, by definition, a global process: we are possibly conscious of anything (and of almost any combination of things). Consequently, any process capable of predicting the content of consciousness needs to be global too, not modular. It would perhaps be more accurate to describe Taylor's theory as an account of our capacity to self-attribute mental states of propositional form. But then it is not clear that attention plays a central role in it, for one could argue that this is just a result of matching prediction of the activity of a module articulating conscious contents, similarly to the account given earlier in 6.1.2 about the possible role of a self-monitoring or metacognitive module in the GW theory.

[35]Fletcher and Frith (2009) argue that many symptoms of schizophrenia can be explained in this vein, i.e. as an undue weighting of certain prediction errors. They suggest that the responsible mechanism behind this is a malfunction of dopaminergic system that encodes precision weighting. Their theory thus elegantly links together the physiological, cognitive and behavioral markers of schizophrenia.

Even though Taylor's theory does not provide a convincing account of transcendental self–consciousness, some of his ideas could gain more credibility if rephrased in terms of the predictive coding theory. I argued that the ability to mark conscious contents as mine cannot be modular. What if we try to explain it in terms of a specific process, rather than a module? Arguably, the process must be more specific than that of matching higher-level predictions to lower-level representations, for that is the fundamental principle of predictive coding and it could not therefore underlie the distinction between conscious and unconscious representations. The solution might be to combine the idea of matching predictions with the crucial role of language in self-reflection.

Looking at the problem from a wider perspective, transcendental self-consciousness is something to us only thanks to our ability to explicitly reflect on our conscious contents. If we were not capable of the latter, transcendental self-consciousness would amount to the same thing as information integration. We can thus hope to shed some light on transcendental self-consciousness (specified functionally as the potential for explicit self-reflection) by investigating the relation between the integration of consciousness and the process of active self-reflection.

Let's assume that self-reflection, or formation of a higher-order thought, involves articulation of the object thought by the language module. Thus the higher-order self-reflective mental state has necessarily a propositional form, while the lower-order state need not have. Articulated thought sustains (predicts) the representational activity of the object-thought[36] - for example, if I observe a complex visual scene, articulating successively what I am seeing will focus my attention on the described features, amplify their representation and that in turn will sustain the activity corresponding to the articulation. Self-reflection could thus be conceived of, at the vehicle level, as a loop of mutually supporting activations between the object mental content and its conceptual representation. The identity of the subject is known non-inferentially for the same reason I know that something is part of my body or that I am raising my hand - by matching predictions from certain sources. Here, the special source would arguably be

[36]For this reason we feel like it is impossible not to think of the pink elephant when somebody asks us to - the higher-order cognitive intention causes tokening of the thought about the pink-elephant.

anything but the auditory system, for articulating information from that source would correspond to a heard utterance. Prediction error signalled by the language module would thus be interpreted as an outer utterance, while low prediction error (high accuracy) of the module with high weighted precision (meaning that the system is attending to conceptualization of its current content, i.e. engaging in self-reflection) would correspond to self-reflective awareness of one's conscious content.

The mechanism proposed here is essentially the same as that proposed earlier in 6.1.2. However, unlike in the previous account, here we don't have a convincing explanation of what renders the self-reflective articulation conscious - for whether it is conscious or not depends, according to the PC theory, on its precision weighting. As far as I can see, there is no principled reason why precision, in the sense of reliability, of some self-monitoring module should vary. If that is the case, precision can vary (and thereby render self-reflective articulation conscious or unconscious) only due to attentional shifts, which leads us to the idea, already expressed earlier, that we engage in self-reflection when it helps to reduce uncertainty (explain away prediction errors) in some other domains.

Naturally, the reason why this mechanism has developed is that we are social animals using language to communicate our ideas. If language had just the simple role of reporting one's current contents of consciousness or otherwise inform other speakers, there would probably be little need to develop the thorough sense of self-awareness. What drives the development of explicit self-awareness (i.e. awareness of the identity of the self in the higher-order state and its object state) is, I think, the fact that we routinely engage in what Sellars called the game of giving and asking for reasons. We are held accountable for our utterances and so we form the concept of a subject of thought as that who is responsible for the produced statements. In social context, this responsible subject is always mapped onto a specific person or body. In its abstract sense, however, it denotes the general idea that every statement presupposes a subject endorsing it. Having this concept formed, one can be explicitly aware of the identity of the subject on this model of a subject endorsing a statement. This is not to say that language is a necessary condition for metacognition in general. There is a number

of metacognitive feats, such as judgements of learning or implicit estimation of one's probability of success, which might have evolved without language. It may well be the case, however, that explicit self-awareness does presuppose language and the recognition of social norms related to its use.

An objection could be raised that saying that language is a necessary condition for self-awareness is trivial because insofar as explicit self-awareness must have a propositional form, that form must come from having language. The point here is not trivial though: language is necessary for self-awareness not just because it serves as the vehicle for our higher-order mental states but because the practice of using it in the community of speakers leads to formation of the key concept thanks to which we can *think* the identity of ourselves, namely the concept of a subject endorsing a statement.

6.2.6 Predictive coding and Kantian echoes

Let me conclude the discussion of the predictive coding theory by noting which Kantian themes it echoes. Of course, Kant's largely epistemological project is orthogonal to the naturalist project of the predictive coding theory (providing a unified account of the place of the mind in the natural world), but that does not preclude some convergence in the positive psychological parts of both theories.

First, Kant's fundamental idea that the world conforms to our knowledge of it because we cannot but experience it using the conceptual structure we have is echoed in the basic idea of the PC theory that we experience the world in virtue of modelling the causes of lower-level neural activity. Perception, in both cases, is a matter of using the pre-existing conceptual structure to explain (synthesize) the manifold of senses. In this rough analogy, intuitions would correspond to prediction errors and lower-level predictions, and concepts would correspond to higher-level predictions. Predictions without prediction errors would be empty in the sense that nothing would drive selection of the best model and hence the conceptual structure, in the form of prior Bayesian models, would be inert - no model could be selected rather than another. Prediction error without predictions would be blind in the sense that nothing would pick up the patterns in the prediction error signal.

Obviously, the important difference is that the PC theory is developed within the broader materialist and evolutionary context whereas transcendental idealism is explicitly developed from an ontologically neutral point of view. As mentioned earlier, however, we can get a coherent and original view of the mind even if we detach the metaphysics of transcendental idealism from Kant's transcendental psychology. It is interesting to note that starting from the materialist view, the PC theory can offer, unlike Kant's theory, an account of why we have the concepts that we do. This is so because in the PC theory the things in themselves (the world as it objectively is, independently of our conceptualization) are not just regulative ideas but instead have a positive explanatory role: their interaction with our body (senses) give rise to patterns of neural activations that can be picked up by the brain, the biological inferential machine.

Another Kantian theme echoed in the PC theory is that we are experientially open only to those patterns in the world that we can pick up using our senses. Kant claims that the a priori forms of intuition delimit the dimensions of our experience. Similarly, the physical properties of neurons and sensory organs delimit the range of effects whose causes the brain then tries to reconstruct using a hierarchy of predictive models.

6.3 Integrated information theory and the dynamic core

In the previous discussion of the global workspace theory and the predictive coding theory we noted that integration of various representations into one conscious state must be ultimately cashed out in terms of causal relations among vehicles of conscious contents (see p. 104). In the global workspace theory, the crucial question is what is the causal mechanism responsible for formation and change of the global workspace - how unconscious representations are recruited to the GW and how conscious ones drop out. A detailed account of that would be the materialist explanation of consciousness according to the GW theory. Similarly, predictive coding needs some account of how predictions from various domain-specific generative models get integrated into one global hypothesis about the state of the world at a time (which is identical to the content of consciousness at a time). I argued that the greatest challenge for both theories is to describe the

information flow among high-level models that are not organized hierarchically (do not converge in a single "master" module). The integrated information theory and the dynamic core, put forward by G. Tononi and G. Edelman, provide currently the most detailed account of the causal structure that underlies the integration unity. I will thus review the integration information theory and try to show how the challenges mentioned above could be met

Tononi and Edelman (1998) propose a theory aspiring to bridge the explanatory gap between the phenomenological structure of consciousness and the informational structure of brain processes. According to their view, the key features of consciousness that need to be explained are integration and differentiation: "conscious experience is integrated (each conscious scene is unified) and at the same time it is highly differentiated (within a short time, one can experience any of a huge number of different conscious states)."[37] When describing what they mean by the unity or integration, they refer to the phenomenological intuition that "each conscious state comprises a single 'scene' that cannot be decomposed into independent components."[38] The integration into one conscious state manifests itself most clearly in phenomena such as binocular rivalry, bistable images (we can entertain only one interpretation at each moment), or the inability to perform two independent tasks that both require conscious processing of inputs. In contrast, split-brain patients, under right experimental conditions, fail to exhibit the integration and consequently lack its benefits as well as constraints.[39] These examples illustrate that the unity is not conceived only as a formal conjunction of features represented at various parts of the brain, but as a holistic state that follows the coherence norm discussed in 3.3.

[37](Tononi and Edelman, 1998, p. 1846)

[38](Tononi and Edelman, 1998, p. 1846)

[39]For example, Sperry (1968) describes a dual task in which a split-brain patient is searching for an item in a pile of test items. Each hand received an object which was then removed and placed in the pile. The subject then searched for the two items using both hands in parallel, each hand searching independently of the target object of the other hand (thus if the left hand grabbed the right hand's target, it would reject it and continued searching). Split-brain patients can thus perform this task faster than normal subjects because the functional disunity allowed for parallel search.

Differentiation is spelled out in terms of information theory as the property of high informativeness of each conscious state. Since there is a vast number of possible conscious states (contents) that we can be in, in response to both internal and external environment, the fact that we are in one and not another state of consciousness is highly informative. The authors illustrate this point by the difference between using a human observer and a photodiode for detecting whether a screen is white or black. Although the photodiode can differentiate a white screen from a black screen as well as a human observer, human observation is more informative because people could react differentially to a vast number of other situation whereas the photodiode has a repertoire of only two states.

Now, differentiation alone is not enough to account for the intensity or breadth of consciousness. To see why, consider a chip with 1000 transistors. The chip can be in a total of 2^{1000} different states. We can imagine a chip with as many transistors as there are neurons in the brain - for example a large photosensitive chip used in digital cameras. According to Tononi (2004), the reason why such a chip would not be conscious, unlike the brain, is that each transistor in the chip is causally independent of the state of other transistors. In other words, no information is integrated by the chip as a whole. For it is the causal dependence of a state of a neuron (whether it is firing or not) on states of other neurons what makes the information borne by the pattern of activation of the presynaptic neurons integrated by the postsynaptic neuron.

Specifically, integration is defined in terms of mutual information among elements of a system: "a subset of distributed elements within a system gives rise to a single, integrated process if, at a given time scale, these elements interact much more strongly among themselves than with the rest of the system."[40] So, what forms an integrated system depends on the power of its parts to influence each other. The authors offer a mathematical formula to represent the degree of integration of a subset of elements and define the concept of functional cluster as the maximally integrated process in a given system. The main idea is that the state of consciousness at any given moment corresponds to the functional cluster

[40](Tononi and Edelman, 1998, p. 1848)

in the brain, which is also called the *dynamic core* (Tononi and Edelman (1998)) or *main complex* (Tononi (2004)).[41]

It should be noted that the concept of information is not related to what we experience as content at the personal level. A system represents information to the extent it enters different states with various frequency, depending on its interaction with the environment. Note that what counts as a state is a matter of theorist's choice of the level of granularity. The brain could be viewed as an information representing system at various levels: at a coarse-grained level, we could model the brain as levels of activation of roughly 50 Brodmann areas, at a fine-grained level each state could be understood as a combination of states of each neuron, and at still a finer level we could define a state as including other cellular and subcellular properties of individual neurons. If the choice of the level of description is arbitrary, then, given that the measure of integration depends on the unit of analysis (how fine-grained the subsets of system S can be), it undermines the idea that the measure of integration maps onto the intensity or breadth of consciousness. Tononi argues that, given the biological properties of neurons, there is a fairly limited spatio-temporal scale at which the brain can be understood as processing information. It is likely that choosing a level of description outside this narrow spatio-temporal window would result in lower measure of integration because, for example, at too short a time interval we could not observe interaction between two neural groups (it takes some time for activation to propagate across neurons) and too large an interval would render the neural groups causally independent. Despite this argument, a perhaps counteintuitive consequence of the proposed measure of integration is that systems composed of units that interact at different spatio-temporal scales could achieve high integration as well. Nothing essentially biological is necessary for consciousness, in this view.[42]

The view is then compatible with (or even inviting) vehicle externalism about conscious contents. Imagine a situation when one is intensely engaging with an

[41] See B for the precise formulation and more details.

[42] "[C]onsciousness should also exist, to varying degrees, at multiple spatial and temporal scales. However, it is likely that, in most systems, there are privileged spatial and temporal scales at which information integration reaches a maximum." (Tononi, 2004, p. 19)

external system capable of undergoing different states depending on the user's actions and thereby, in turn, elicit different states in the user (think of a virtual environment in a computer game, for example). Should the mutual information between the user and the system in the environment be very complex, it could, in principle, happen that the complex with the highest amount of integrated information involve the external system. It could be argued that formation of such a complex that includes an external system is unlikely because either the external system and the brain are likely to work at different temporal scales (hence generating low mutual information due to different sensitivity to change in time), or, even if they worked at the same temporal scale, the temporal rate at which information is transduced by sensory organs is too slow.[43] Nevertheless, this is a physical constraint that could in principle be overcome by technology.

Tononi (2004) uses examples of simple networks to show it is unlikely that the whole neural network would form the functional cluster with maximal integrated information. Intuitively, including specialized subsystems that process information relatively independently and only pass the result to the rest of the network would result in a decrease of effective information.[44] In the discussion of the global workspace theory, we noted that the informal criterion of recruiting a neural group into the global workspace is the relevance of its representation. Integrated information theory can be interpreted as showing that relevance can be specified as effective information between that neural group and the rest of the system.[45] If we look at the brain as an isolated system (as the information integration theory does), then effective information among neural groups may

[43]This is known as the bandwidth problem that is often faced in engineering a brain-device interface. Clark (2009) uses the bandwidth problem to argue for internalism about vehicles of conscious contents.

[44]'Effective information' is a technical term which is supposed measure the degree of mutual influence that two areas have on one another. See section B for its precise formulation.

[45]Strictly speaking, the reason for "recruiting" a neural group into the global workspace would not be the effective information between it and the current GW but rather the fact that the union of the neural group and the subsystem that currently forms the GW will have a higher measure of information integration than the subsystem alone. However, such a situation can occur only if the effective information between that neural group and the current GW is high, relative to other candidates competing for access to the GW.

seem to capture very little of the sense of ecological relevance used in the GW theory. However, since informational (causal) relations among neural groups are formed over time thanks to interaction with the environment and various neural learning mechanisms, we can say that parts of the neural system have informational relations they have *because* these proved to be ecologically relevant. One could then argue that effective information corresponds to relevance (as long as one accepts this naturalist reduction of relevance in terms of adaptiveness).

An important feature of networks with high amount of integrated information is that they combine both functional specialization and integration. Intuitively, if every neuron were connected to every other, then no information would be integrated because every neuron would pass information to every other. On the other hand, if the brain consisted only of specialized and largely independent areas (high functional specialization), not much of information would be integrated within the whole network. The brain thus needs to strike the right balance between specialization and integration. That this is the case seems to be empirically supported by studies of brain's network properties (both functional and structural connectivity obtained by fMRI and DTI respectively). Network properties desirable for optimal information flow across the brain seem to be linked to some cognitive capacities, for example attention and its deficits (ADHD), or fluid intelligence. Another line of supporting evidence comes from theoretical research of criticality.[46] Shew and Plenz (2013) review theoretical accounts of criticality in the cortex, showing that when the brain operates at criticality, it maximizes number of information-theoretic measures such as dynamic range (the amount of stimulus features distinguished by a differentiated activation pattern), information transmission (mutual information between stimulus and the brain

[46]Criticality refers to the rate with which activation spreads in the brain. The critical rate σ=1 means that each firing neuron would cause the firing of, on average, one other neuron. Supracritical values of $\sigma > 1$ would lead to an undesirable dynamic in which the whole brain would be flooded with activation after few iterations (which happens during epileptic seizures, for example), and subcritical values would mean that an inititial activation (e.g. from a stimulus presentation) dies out soon without making any effect. Criticality thus depends on the right balance between excitatory and inhibitory signalling. For a more detailed overview of the concept of criticality and its empirical validation, see Shew and Plenz (2013); Arviv et al. (2015).

state), and information capacity (entropy). Theoretical predictions that the brain should operate at criticality have recently been confirmed empirically by Arviv et al. (2015). Although these studies do not bear any direct relevance for consciousness, findings about functional benefits (manifested at the behavioral level) of these information-theoretic measures render information theory relevant for the study of consciousness in general, and the information integration theory plausible in particular.

In agreement with the global workspace theory, the information integration theory also implies that the dynamic core (i.e. the functional cluster of areas with high degree of mutual influence) may also change over time provided that the capacity of areas to influence each other's states is dynamic. However, a change in the content of consciousness is here understood as a transition from one state of the dynamic core to another, while the composition of the dynamic core may remain unchanged. Examples of change in dynamic core would be changes in levels of consciousness (e.g. deep sleep or coma would be a state of the brain where the dynamic core has relatively low measure of information integration, compared to being awake) or changes in functional connectivity, such as the transition from the resting state to a task-oriented state (switching from the default mode network as the dynamic core to the task-positive network).

To support the theory, Tononi (2004) offers extensive discussion that relates informational properties of the dynamic core to phenomenological characteristics of consciousness and to the neural underpinnings that possibly realize this informational structure. For example, the theory offers an explanation of why it takes more than 100-200 ms of sustained neural activity to produce conscious sensation of a stimulus and why, on the other hand, we are able to react unconsciously yet differentially to subliminal stimuli. The authors go as far as to say that the IIT provides a framework for thinking about qualia in informational terms and can account, for example, for their irreducibility and interdependence (what it is like to see red depends on there being potentially something it is like to see blue).

The theory explains why activation of a neural group that arguably represents a feature may not contribute to conscious content. If the neural group is outside the dynamic core, it may still represent the feature unconsciously and

drive differential automatic response - it just is not informative enough for the whole system. This explanation is thus very similar to that provided by the GW theory. However, a rather strange implication of the IIT is that the *actual* degree of consciousness depends only on the *potential* of information states to influence each other. Tononi (2004) explicitly admits this seeming paradox but adds that it is quite common to describe some natural properties in terms of dispositions, e.g. mass as the disposition to attract bodies. He concludes that

> [I]n this view consciousness corresponds to the potential of an integrated system to enter a large number of states by way of causal interactions within it, experience is present as long as such potential is present, whether or not the system's elements are activated. Thus, the theory predicts that a brain where no neurons were activated, but were kept ready to respond in a differentiated manner to different perturbations, would be conscious (perhaps that nothing was going on). (Tononi, 2004, pp. 19-20)

Importantly, the integration information theory avoids the category error of identifying properties of local neural groups (e.g. activation) with consciousness of a specific content. Level-consciousness[47] is a property of a system, not of a state.

As Edelman (2003) points out, the ontological framework that befits the dynamic core theory the most is probably epiphenomenalism since the theory still holds that it is actual brain states which cause our actions, and that consciousness is a higher-order property of brain processes and is not, as such, causally efficacious. We could argue that the theory is also congruent with property dualism or even panpsychism, since its numerical measures of integration and differentiation in terms of information theory permit ascribing some, albeit minimal, degree of these measures to any system (biological or mechanical) and there seems to be no principled reason to set a specific threshold in these measures for consciousness to arise. Second, the dynamic core theory is clearly a functionalist one: any

[47]Both authors fail to use the distinction between level-consciousness and content-consciousness but their theoretical disccusion makes it largely clear that the proposed measure of information integration is a measure of level-consciousness and that the conscious content corresponds to the actual state of the dynamic core.

system whose states have the appropriate informational properties would satisfy the criteria for consciousness. It is thus possible that galaxies, populations or ant colonies are systems with great enough integration and differentiation that they are conscious (although at different spatio-temporal scale than we are). Interesting as these metaphysical implications may be, they are not directly relevant to the unity of consciousness.

6.3.1 Contrasts and comparisons of the IIT with the GW and PC theories

The information integration theory regards the unity as a defining feature of consciousness whereas the global neuronal workspace theory regards the unity as an emergent property of the specific neural architecture which realizes the global workspace. However, the difference could be just in emphasis: it is plausible that the architecture which realizes the global workspace in the brain is necessarily such that it also yields high level of integration as defined by the IIT.

The theoretic concept of information integration is currently the most specific account of what the integration unity of consciousness amounts to at the vehicle level. To compare, the GW theory understands the unity as global availability across a variety of modules which constitute the global workspace at a time, but does not specify the mechanism responsible for formation of the global workspace beyond saying that a module is recruited in the global workspace as long as there is a loop of sustained activation between it and other workspace modules. But why is it that these modules, rather than others, currently constitute the global workspace in the first place, i.e. why do these modules causally interact in the particular way that is held to be necessary for consciousness? At this point, Baars and Dehaene defer the explanation to notions of context and relevance: modules are recruited in the workspace because their representation is currently relevant with respect to the context the agent finds herself in. The IIT specifies relevance as effective information: a representation (a pattern of activation in some neural group) is conscious only if it makes a difference in the rest of the dynamic core (i.e. if changing the state of that neural group would change the state of the dynamic core). Importantly, understanding this kind of causal interaction

as *relevance* makes sense if we incorporate in the picture an adaptive learning mechanism that effecively shapes the state space of the neural network based on the organism's interaction with the environment. This part of the picture is well described by the predictive coding theory because it argues from the start that the organism minimizes prediction error only if the generative models learn to predict the activation of sensory neurons which, in turn, reflects causal patterns in the environment.

An important difference between the GW theory and the IIT is that while the GW theory implies that neural activation is necessary but not sufficient for conscious representation,[48] the IIT denies that it is even necessary. According to the latter, it is necessary that neural groups are connected to other groups so that they can exert influence on them. But it is not necessary that the influence is *actually* realized. A representation corresponding to a state of no or minimal neural activity in a particular neural group may be conscious provided that changes to it would produce changes in the rest of the system.[49] Clearly, given that the brain activity is controlled by both excitatory and inhibitory neurons, zero activation in one neural group (containing, among others, inhibitory neurons) may cause disinhibition of a connected neural group and thus change its state. To use a simple but unrealistic example, suppose that exposure to complete silence causes neurons in the primary auditory cortext to be silent. Phenomenologically speaking, we can be acutaly aware of the silence, e.g. when we find the silence suspicious and try to prick up our ears. In the IIT, this is so because the whole neural network is poised to be influenced by any perturbation to the auditory cortex. In the PC theory, we would be conscious of the silence because we have

[48] "To enter consciousness, it is not sufficient for a process to have on-going activity; this activity must also be amplified and maintained over a sufficient duration for it to become accessible to multiple other processes." (Dehaene and Naccache, 2001, p. 14)

[49] The claim that a neural group with zero activation could still represent something is pushing the concept of neural representation to its limits because it is then not clear how we could ever reliably establish what a particular neural group represents. Although we could measure effective information between areas using data from brain imaging, the information-theoretic relational concept of effective information does not allow us to tell what the *relata* represent. It is likely, I think, that the IIT is committed to such a holistic conception of how the brain represents that it undermines the concept of neural representation.

a strong prior for at least a low level of noise heard, hence *some* activity of the auditory cortex would be predicted and a precise prediction error generated.

6.3.2 The subject unity and the dynamic core

The subject unity of consciousness is not addressed by the IIT, although Tononi and Edelman (1998) claim that the dynamic core *is* the subject, in the sense of being a point of view where information is integrated. This identification, however, seems to be philosophically naïve. Any system composed of interacting parts may have a main complex (dynamic core) integrating some amount of information - if we accepted that each such complex constitutes a subject, the concept of subject becomes too vague to capture what seems to be essential about human consciousness, namely the capacity to be aware of one's state of consciousness as such. Furthermore, the theoretical specification of the dynamic core cannot offer a convincing explanation of why we should understand the dynamic core as an instantiation of perspectival self-consciousness either - unless the theory is backed by an account showing that the reason why effective information among neural groups that are part of the dynamic core is high *because* they all represent things from a common perspective.

How could then the information integration theory account for self-consciousness? The theory implies that we will be metacognitively aware of our mental states only if that awareness is somehow informative (relevant) to the rest of what one is conscious of at the moment. For the sake of illustration, let's assume that a subset A of system S has the capacity to make higher-order representations of the states of S-A - the complement of A to S (whatever the higher-order representation might be). Thus each possible state of A would be a higher-order representation of some states of S-A (given that A is smaller than S-A, then A can differentiate only between some states of S-A, not all of them). Effective information from S-A to A is consequently relatively high - in an ideal case it approaches maximal entropy possible for A (see B for details), as that is when A uses its resources to differentiate maximum possible states of S-A. However, effective information in the opposite direction (from A to S-A) is high only to the extent to which the higher-order representation can inform the rest of the

system. Arguably, for A to be part of the dynamic core (and hence for us to be reflectively aware of our conscious state) the effective information must be relatively high in both directions. As a consequence, we become conscious of the always-ready metarepresentation only if it makes a significant difference to the rest of the system. Intuitively, we employ metarepresentation in cases like planning (how could a situation unfold), dealing with uncertainty (e.g. revising one's calculations) or moral reasoning (e.g. reflecting on what are the true intentions of myself and others before making a decision).

Since the higher-order representation would be informative to the rest of the system in virtue of the content, why is the higher-order representation needed when the content is already there, so to speak? Could not the rest of the system be informed by the original lower-order representation? The information integration theory shows why that would be inefficient: in order for the lower-order representations to possibly inform the rest of the system, the corresponding areas would have to have diffuse connections to the whole brain which would in turn reduce functional specialization and hence the amount of integrated information. The same reasoning also explains why it is unlikely that there would be only one subsystem for higher-order representing. A network containing areas capable of progressively more abstract higher-order representations of lower-order features will integrate more information than a network with a single centre dedicated to higher-order representation of the rest of the system (if that is even possible). It might well be the case that linguistic representation of conscious contents has the widest and most abstract metarepresentational scope, but many metacognitive tasks can be achieved without recruiting this specific kind of metarepresentation.

Regarding self-reference without identification, the theory does not offer much of an explanation, mainly because the theory is concerned more with level-consciousness and less with actual contents of consciousness. Given that the actual state of the dynamic core corresponds to what we are conscious of, then any conscious content, which is identical to a state of a subset of the dynamic core, is trivially a part of consciousness. The explicit higher-order recognition of some lower-order state as mine is a metarepresentation that would be included in the dynamic core only if it were informative to the rest of the system, e.g. when

a subjectis asked to report her thoughts. The implanted thoughts phenomenon could be correspondingly described as a case in which both the delusional object thought and its metacognitive attribution to someone else are part of the dynamic core. The IIT does not explain why the metacognitive attribution is wrong. Nonetheless, the theory seems compatible with the explanation based on expectation matching described in 6.2.

6.4 Summary

In this chapter I reviewed three influential theories that provide partial explanation of the unity of consciousness at the neural level. Generally, the theories are compatible with each other and each is suited to explain some aspects of the unity better than others. If the unity of consciousness can be analyzed into different components, as Hurley claims, for example, we can hope to provide a satisfactory naturalistic explanation of the unity in a piecemeal fashion without having to provide a grand unified theory of how the brain instantiates consciousness. Certainly, an explanation of the unity that is based on selecting ideas from different theories will be satisfactory only to the extent the theories have roughly the same understanding of what consciousness is. Since consciousness is such an elusive concept that even philosophers, despite agreeing that it is a real thing, have not agreed on its essence, we should not expect it from theoretical neuroscientists either. What we may expect, however, is a growing convergence among empirical theories of consciousness. This convergence would support the claim that consciousness can be naturalized. This, I think, is the case. The three theories reviewed in this chapter exemplify such a convergence. So let me summarize the most important ideas that together could set the ground for naturalizing the unity of consciousness.

The global workspace theory was probably the first to account for consciousness in terms of dynamic relationships among unconscious modular processors. Baars's initial model using an information processing diagram ("boxology") was later fleshed out in terms of a dynamic pattern of neural amplification among areas by Dehaene and others. The global workspace is constituted by areas that interact with each other more than with the rest of the system. The set of ar-

eas that constitute the global workspace at a time is maintained by loops of sustained activation - that is, by a pattern of causal interaction among areas that is relatively stable and self-supporting. Since the theory identifies the GW with consciousness, the content of consciousness at a time corresponds to what the areas recruited in the GW represent, and as a consequence, the integration unity corresponds to the pattern of mutually supporting local activations. As mentioned earlier (3.2), the identification of neural activations with particular representations is notoriously problematic. With that in mind, a charitable interpretation of the GW theory would be that insofar as we can decompose the content of consciousness at a time into constitutive representations (now taken from either auto- or heterophenomenological point of view), their integration into one conscious state is a matter of causal relations among the corresponding neural activations. More specifically, the causal relations are such that each neural representation supports, and is supported by, the other conscious representations. Since the pattern of such mutually supporting activations is what renders the corresponding representations *conscious*, the unity thus defined is constitutive of consciousness.

It can be stated, I think, that patterns of local neural activations support each other in virtue of the content they represent. Naturally, at the subpersonal level of description the patterns in question enter the loop of sustained activation in virtue of the physical factors governing signal propagation in neural tissue - network of axonal projections, synaptic weighting, modulation, spiking frequency, etc. Nevertheless, it can be argued that the brain exhibits these particular properties so that they support some particular behavioral patterns - interactions with the environment - that actually constitute the content. The range of differential response to stimuli is mostly learnt during ontogeny, and learning at the physical level is a matter of creating new networks and adjusting the neural factors that influence signal propagation in the brain. (To be more precise, we are surely capable of complex behavior that is contenful but cannot be described simply as differential reaction to some stimulus feature, for example understanding humour or feeling lucky. In such cases it is perhaps unlikely that there would be a localized neural activation corresponding to the content “humorous” or “being lucky”. But

even if such representation is distributed over a large part of the cortex, it is safe to say that it is causally related to (has connections to) other neural representations (e.g. the action of smiling) in virtue of their content.)[50]

The idea that neural activations form the global workspace in virtue of the represented content is an important part of the explanation of the higher-order awareness of one's conscious states that I put forward above. The core idea is that the neural activation representing content of one's consciousness supports (and is supported by) by the activation corresponding to that content because the former *describes* (correctly) the latter. However, the correct description itself is not sufficient for the metarepresentation to enter the GW, otherwise we would be constantly engaged in this self-monitoring activity. The metarepresentation enters the GW (we become conscious of our lower-order conscious contents as objects of consciousness) because it is relevant to the other representations currently recruited in the GW, including goals and actions. For example, I may find myself coming to the kitchen and realizing that I forgot why I went there and consequently start reflecting on my intentions and thoughts.

Again, the conceptual problems related to identifying representations and metarepresentations in the brain appear, this time more strongly because the activation underlying metarepresentation is in principle the same as that of ordinary representation, while at the conceptual level there is a clear distinction between the metarepresentation and the object representation. Still, this problem is not decisive, although it implies, as I argued, a blurred distinction between metarepresentation and any higher-order, abstract representation based on lower-order features.

[50]Representation is here understood broadly so as to include non-descriptive mental states such as motivations and actions. For example, neural representation of a movement corresponds to activations of neurons at premotor and motor cortices that intitiate the right sequence of muscle actions. How far the neural representation of a movement goes depends on how specifically the movement is conceived. For example, if it's just raising an arm, the relevant activation might be limited to premotor cortex; if it is raising one's left arm along some specific trajectory, the representation would probably include cerebellum. The latter, however, would not qualify as a representation in the strong sense defined in 3.2.

Given that content plays a significant role in shaping relations among neural representations, we can understand why contents that we are conscious of at a time are largely coherent. Children learn progressively more abstract patterns of relations among represented features of the environment. Eventually, the acquired connections among representations will be such that a representation of one fact will inhibit inconsistent representations. For example, until a child learns basic folk-physical laws of solid objects, she is not surprised by seeing a ball passing through a wall, and would arguably represent the ball as being the same object throughout its trajectory. After she learns this abstract feature of solid objects (i.e. forms a strong set of prior expectations about how solid objects interact), she will be surprised and eventually come to represent the passing ball as different before and after it apparently passes through a wall - because by then that will be the more plausible interpretation of the cause of her sensations. The coherence norm manifested in consciousness can thus be seen as a result of developing a cognitive model of the world.[51] The pressure to develop a coherent model of the world is inherent to all self-organizing systems that seek to maintain a stable internal state despite changing external conditions.

Explaining cognition and action in terms of predictive models of the world is the goal of the predictive coding theory. The theory does not primarily concern consciousness though - it shows how cognition and intelligent action naturally arise in living things, following general physical and biological laws. Although the account of consciousness within the predictive coding theory is relatively marginal, the theory brings two main contributions to the consciousness science: 1) it offers a philosophically robust account of representation, and 2) it provides a unified account of cognition and action that is both parsimonious (using only three basic elements of prediction, prediction error and precision weighting) and versatile (a wide range of mental phenomena can be explained by a hierarchical mechanism using these three principles).

[51]To be clear, this constraint is not normative in the strong sense of following from a set of rules that can be articulated and enforced at will. If consciousness is subject to the coherence norm in this strong sense as well (as the psychoanalytic theory holds, for example), it is because a significant part of the world that human beings need to make sense of is the social and linguistic world, which do follow conventional rules.

Regarding the first point, by saying that the acount of representation is philosophically robust I refer to these features: 1) representation, understood as prediction formed by a generative model, owes its existence to the history of organism's interaction with its environment. The Bayesian update rule and the proposed architecture show how a deterministic system with no initial knowledge can learn to represent patterns in the environment - which is a part of what it takes to have contentful states. 2) There is no difference between representation-for-action and representation-for-perception; representation is not passive information that leads to behavior by being further interpreted and then acted upon (cf. the sharp distinction between input processing and executive functions in classical AI). 3) Representation is holistic in the sense that predictions of some generative model represent what they do in virtue of the model's connections to other models. The first feature is crucial in that it specifies the causal mechanism by which organisms acquire contentful states. The latter two points render the account robust to the objections often raised against the representational or computational theory of mind.[52]

When it comes to explaining the unity of consciousness in particular, the predictive coding theory is the most promising thanks to being congenial in many respects to the Kantian view of the mind in which the unity is a pivotal concept. Two such similarities are worth repeating. First, Kantian concepts are principles of unity, i.e. functions that combine the manifold of intuitions into one cognitive unit that can be subsequently combined with other concepts in a systematic way, e.g. into a proposition. In a similar vein, generative models "combine" lower-level sense-data (prediction errors) into a prediction (hyptohesis about the cause of the lower-level error) which can, in turn, be combined by models further in the hierarchy. Second, Kant holds that apperception is ultimately unified in the conception of the objective world. That is, our concepts and cognitions display the systematicity and combinability they do so that they help us make sense of the world as objective and us as parts of the world. Similarly, the predictive coding theory explains that representations can be combined so that the organism is able

[52]For a detailed discussion of representation in the predictive coding theory, see Gładziejewski (2016).

to navigate the world using maximum information at its disposal. The reason why all information is integrated in one point of view (hence the reason why there is no functional disunity, safe for pathological cases) is because cognition serves to control one organism, or one vehicle of genes - if we want to emphasize the selective advantage of unified cognition in the evolutionary perspective.[53]

The predictive coding theory accounts for generating predictions (the best guess of the current state of affairs) in terms of a hierarchical architecture in which a higher-level area predicts the activity at the lower-level area and receives prediction errors as its corrective feedback. The hierarchy, however, cannot go all the way to the top, converging in a single all-encompassing predictor, for that would be an equivalent of the Cartesian theatre. I argued that at the level of higher cognitions (e.g. understanding causal relations, people's intentions, etc.), predictive models are likely to be organized in a network with reciprocal connections the strengths of which reflect the ecological importance of inferences based on associations of their respective representations. The consequence is that consciousness may be unified only partially.[54] To illustrate partial unity, consider again an example from playing a racket sport such as badminton. Our estimate of the shuttle's trajectory and our motor action may be based on integration of visual information (seeing the shuttle move), past experience (observed patterns of play, e.g. playing a cross-court lift after straight netshot), and our own body schema (where and how we are positioned and what movements need to be made). In estimating the trajectory, an inexperienced player may fail to integrate the sound of opponent's racket hitting the shuttle. However, the sound is likely to be integrated with seeing the racket hitting, in the sense that the seen racket is recognized as the source of the sound. The partial unity is thus exemplified by the fact that the trajectory estimate is integrated with one's movement, expectations

[53]To continue with this speculative listing of similarities, we can argue that Kant's emphasis on the mind's activity in making sense of the world (as opposed to passive feed-forward processing of information) fits nicely with the main principle of the PC theory that all cognition is a matter of matching its best guess about the world against incoming sense-data.

[54]Partial unity of consciousness means that for conscious representations A, B, C, the relation 'being unified/integrated with' may not be transitive: A can be unified with B, and B with C, but not A with C. See Hurley (2003) for a detailed view.

and vision, but not with hearing - although vision and hearing are integrated together as manifested by binding the sound to the seen racket.

The importance of partial unity is that when we say that the predictive coding architecture enables us to represent the objective world, it needs to be kept in mind that the world is never "predicted" in its totality - our sense-making of the world at a time is always limited to that section which interests us the most. So, we may be attributed with having the idea of the objective world (as a whole) only in the sense in which our whole predictive machinery is set up to represent the widest possible range of states of the world. At a time, however, we are able to make sense only of a part of the world state that we are disposed to represent.

There are two reasons why we might tend to think that our consciousness at a time is completely unified. The first is that we often fail to notice regular shifts in attention whereby we sample the world (or internal states) for useful information.[55] The second is that we might be quite used to articulating the content of consciousness in language. If we understand this explicit, articulated self-reflection on the model of perception, where we focus the linguistic spotlight, so to speak, on a part of our conscious state in order to articulate it, we are naturally led to conceive of the conscious state as a fixed thing, at least for the time it takes to articulate the conscious content. And since language and its use incorporates logical and semantic relations, we may be led to think that by tokening an articulated reflective state we also token logical consequenes and presuppositions of the statement that conveys our current content of consciousness - and therefore that we are conscious of them. That this is not the case is illustrated, for example, by the experimental finding, described earlier in 3.3, in which seeing a die for a short time and being aware of it leads us to think that we have also been conscious of a specific number of dots on its side.

[55] An example of external sampling is saccadic movements that yield a stable conscious percept. A phenomenologically conspicuous example of internal sampling would be assessment of a character of a person by sampling one's long-term memory for observed behavior of that person. In general, the predictive theory holds that sampling is the brain's way to estimate the prior probability distribution. It is an unconscious process characterized purely in computational terms - the examples above should be understood as analogies rather than instances of sampling in the technical sense.

Since we become proficient in reporting our conscious states as we grow up, we rarely find ourselves not "knowing" what we are conscious of; and this produces the experience that no matter where we turn the linguistic spotlight (self-reflective focus), there is always something there. Given that on the perceptual model of self-reflection the object conscious state is considered to be stable, the natural conclusion is that our conscious state is completely unified. If, on the other hand, we reject the perceptual model of self-reflection and assume that changes in self-reflective focus go hand in hand with changes in the object conscious state, we are led to the conclusion that the apparently complete unity is an artifact of systematicity of language - of the unity of language in virtue of its systematicity and inferential relations among its elements.

Articulating conscious contents in language is crucial for the subject unity of consciousness. I argued that the basic mechanism of matching predictions to the actual lower-level states may explain the sense of ownership of thoughts (Shoemaker's self-reference without identification, i.e. recognizing my thoughts as *mine*, non-inferentially) in a similar way as it is used to explain the sense of agency and body ownership. The sense of ownership of thoughts (and by extension of perceptions, intentions, etc.) depends on the learned ability to report on conscious contents overtly. To put it simply, we recognize our thoughts as ours when the prediction error resulting from matching the 'conceptual' representation of conscious state to the actual conscious state is low. However, this picture is problematic in that the 'conceptual' model would need to be able to predict (represent) *any* conscious state, hence it cannot be domain specific. A system capable of representing another system must be at least as complex as the object system itself (to the extent the representation is complex, as opposed to being simple or selective).

This problem with a 'global' metarepresentational module will appear less severe if we accept the quite plausible view that linguistic metarepresentation is more abstract (less specific) than the object conscious state it represents, which is richer in domain-specific information The metarepresentational module could thus still be modular, provided that it is widely connected to many centres. This may still invite the objection that if the capacity for conceptual metarepresenta-

tion is modular, we should find cases in which a localized damage in the brain leads to inability to explicitly think about and report contents of consciousness without affecting most of the subject's conscious behavior (except for those activities that benefit from attentional control due to metarepresentation/explicit metacognition). To my knowledge, there are no such cases. However, we can just reject the assumption that metarepresentational capacity is neatly localized in the brain. Since the capacity to use language involves many cortical areas (not just Broca's and Wernicke's areas as historically thought), it makes a good sense to assume that the metarepresentational capacity is not localized either.[56]

Finally, the discussion of the integrated information theory showed how the integration unity can be specified in terms of concrete information-theoretic measures. The actual specification of the measure of information integration is not as important as showing that an information-theoretic approach can provide a plausible account of two crucial features of consciousness: integration and differentiation. Given the evidence that various information-theoretic measures of the brain map onto behavioral differences, we have good reasons to believe that applying the information-theoretic framework to the unity of consciousness provides us with more than just untestable speculations. The claim that informational relations are the right level of description when trying to explain the unity of consciousness at the vehicle level can be further backed by noting that the other

[56]Note that nothing in the predictive coding theory implies that prediction modules must be localized in smaller areas of the brain. The hierarchical architecture is primarily an architecture of abstract informational relations (see, for example, (Pezzulo et al., 2015, fig. 6)). Although connections conveying predictions and prediction errors must by physical, that does not necessitate that the neurons responsible for predicting activity at the lower level must be physically close to each other. Certainly, looking at the simplified hierarchical structure it would make a good ecological sense for the brain to have the prediction modules localized, in order to minimize the total length of axonal projections and hence maximize the speed of signal transmission; and this seems to be the case for low-level sensory processing that appears to be confined to localized sensory areas. For higher level predictions that code increasingly more abstract features of the world, the prediction areas would need to be connected to many contributing areas all over the cortex, hence the brain would not economize on total connection length by putting the "prediction neurons" together anyway - doing so would actually by maladaptive as it would make the brain more likely to suffer from a loss of high-level function in the case of localized damage.

two theories presented here, namely the global worskpace theory and the predictive coding theory, can both be interpreted along the information-theoretic lines, especially regarding the unity as such.

Using the notion of information in philosophical discussion of consciousness has often been met with suspicion because of its association with classical AI and the computer metaphor for the mind. It is thus important to emphasize that information, as it is used in the PC theory and the IIT, is not regarded as an interpretable symbol.[57] The informational states assumed by the IIT and the PCT need not bear any content recognizable at the personal level. However, the theories hold that content is realized by patterns of (transitions of) informational states in relation to states in the environment.

To summarize the assessment of the three theories, I hope to have showed that the theories: 1) are to large extent compatible and can thus be interpreted as complementing each other in their respective areas of focus; 2) discuss consciousness mostly at the computational level of description with little specification and supporting empirical evidence at the level of neural implementation; 3) try to account primarily for the integration unity - a fitting account of self-consciosuness and the subject unity requires a considerable extension of the theories; 4) the predictive coding theory is the most promising theoretical framework in contemporary neuroscience to account for the unity of consciousness because many aspects of the unity recognized by Kant and Hurley can be most readily described in its terms.

[57]Cf. Searle (1990) and his Chinese room argument.

7 Conclusion

> Why is the unity of consciousness so mysterious that we get ourselves into embarrassing positions trying to understand it? The reason has something to do with the way the unity of consciousness seems to hover between the subjective and the objective realms, to distance itself from each in turn, and to have both personal and subpersonal aspect. (Hurley, 1998, p. 41)

Hurley's quote provides a succinct diagnosis of the difficulties faced by anyone who tries to explain the unity of consciousness.

I started this work by describing the various meanings associated with the concept of the unity of consciousness. I narrowed down the concept of the unity pursued in this work to what I call, along with Bayne and Chalmers (2003), the subject unity and the integration unity, both considered as unity of conscious experience at a time (synchronic). Building on the interpretation of Kant in chapter 4, I argued that these two kinds or views of the unity of consciousness cannot be treated separately. This brought self-consciousness into the picture and the notion of the unity of consciousness thus gained greater conceptual complexity. Let me thus first summarize the key points of the conceptual analysis.

First, we can make sense of experience only insofar it can always be accompanied by the 'I think'. Without this transcendental self-consciousness, things would be "represented in me" without them being something *for me*. This is to say that our notion of conscious experience entails self-consciousness in the sense of being conscious of the representations as mine. Without it, we could perhaps speak of representations from the third-person point of view and admit that they guide behavior, but we would not call them experience, as we don't call computer representations experience. The transcendental self-consciousness involves com-

bining representations into complex thoughts and being non-inferentially aware of oneself as the common subject of the individual representations. Transcendental self-consciousness is to be contrasted with empirical self-consciousness or inner sense - the attribution of properties to myself as an object in the world. In 5.2 I tried to clarify transcendental self-consciousness using Shoemaker's account of self-reference without identification. The key point is that the unity of consciousness does not depend on a subject assembling representations in his consciousness or making judgements of its identity across multiple representations - such view would lead to the infinite regress known as the homuncular fallacy. Rather, transcendental self-consciousness is a feature that comes with the way contents are unified (synthetised) in the perspective of a single agent.

The reference to agency is crucial here: there would be no reason for a completely passive system to be self-conscious. The contribution of agency to self-consciousness is that acting succesfully (meeting one's goals) fosters formation of egocentric, perspectival representations - representations of which (external) states of affairs will bring about *my* desirable internal states and which of *my* possible actions will bring about those states of affairs. Meaningful agency requires feedback from what is acted upon to the agent; and this feedback loop centered around the agent promotes formation of representation of oneself as distinct from the world. Failures to achieve something often motivate learning that the world is not the way *we* thought it was (error in perception) or that *our* action did not bring about the intended effect. Now, since Kant built up his transcendental psychology from the perspective of transcendental idealism, he had to formulate this point in terms of the spontaneity of the mind, not of agency of an organism as a material thing. Nonetheless, the argument is similar in that activity (agency) is held to be constitutive of the unity in virtue of founding the distinction between an active subject and the world as the source of that to which the agent is passive.

This point can be nicely illustrated in the framework of embodied cognition. Since one's actions often directly influence one's perceptions (think of a movement and its predictable impact on what the agent sees) and vice versa, such interactions with the environment motivate formation of primarily *egocentric* representations - what will be the impact of this action on *my* internal states, what

action to take to satisfy *my* goals depending on *my* perceptions, etc. Granted, at a low level of complexity of an agent's interaction it may be unnecessary to talk about representations (let alone conceptual ones) to explain them, but even then the agent's behavior would arguably be sustained by egocentric action-perception feedback loops. For the feedback loops are what constitutes the invariant of intention despite sensorimotor contingencies. Their presence allows us to adopt the intentional stance towards that agent[1] and forms the basic ground of self-consciousness that further requires complex representational capacities. Kant's doctrine of syntheses is an elaborate attempt to specify what exactly these complex representational capacities involve.

The doctrine of syntheses specifies the kinds of activity the mind does to support the experience of the world. The syntheses can be understood as processes that bind together sensory discriminations and other sources of information in order to yield a coherent and ecologically useful representation of the world, and enable one to recognize herself as the common subject of the constitutive representational elements. These syntheses can be well described in terms familiar to cognitive science. The main conceptual challenge is thus to account for the relation between the synthetic activity and transcendental self-consciousness. For this purpose I reviewed Shoemaker's account of self-reference without identification and Hurley's thorough analysis of the conceptual pitfalls present in discussions of the unity of consciousness.

The discussion of Hurley's work clarified that the unity of consciousness needs to be explained at the subjective as well as objective level, and that the objective part of the explanation needs to specify causal relations among vehicles of conscious contents. Hurley's just-more-content argument concludes that purely subjective account of the unity, that is one that accounts for the unity in terms of conscious contents, is impossible (see 3.3 for details). Conscious representation of oneself as the common subject of various representations *presupposes* their unity, it cannot constitute it. Next, founding the unity in coherence among contents is not sufficient because it is possible for a set of coherent contents to be realized (tokened) by distinct consciousnesses. Some degree of coherence of conscious

[1]Or 'system', to avoid the implicit assumption of intentionality inherent in the concept of an agent.

contents is necessary for the unity, but it alone cannot explain what it takes to have a unified conscious representation.

What could be the objective account of the unity then? Previous arguments showed that the objective account must concern tokens of conscious contents. More specifically, the unity needs to be explained in terms of causal relations among their vehicles; coherence and the possibility of self-reflection manifested at the content level ought to follow from it. Clearly, devising such an account is possible only within the naturalistic view of the mind that is in stark contrast with the position of transcendental idealism from which Kant analyzed the mind. Assuming vehicle internalism and identifying vehicles of conscious contents with brain states, devising such an account is the proper domain of cognitive neuroscience. It needs to be emphasized that not any unity found at the vehicle level needs to be *the* crucial objective component or mechanism of the unity of consciousness - be it neural firing synchrony or activity at a specific place in the brain. That a particular process or mechanism underlies the unity of consciousness should not be assumed only on the ground that the process or mechanism is a common factor of all conscious representations. Rather, the assumption is justified only to the extent the proposed mechanism fulfills the function of (unified) consciousness. As a consequence, there may in principle be more than one type of process or mechanism involved that fullfills the function.[2]

[2]Despite the multiple realizability *in principle,* it is still plausible that the range of mechanisms or processes underlying the unity at the vehicle level is fairly limited given our physiology and the temporal scale at which we need to act and therefore integrate information. Nevertheless, the challenge raised by vehicle externalists shows that under certain conditions it makes sense to assume various mechanisms of the unity even in cases of individual people. Consider split-brain and acallosal patients that rely on cross-cuing to forward information to the other hemisphere. Where this cross-cuing is automatic and unconscious, it is a mechanism that uses external vehicles to support a unified conscious representation. (If the cross-cuing were conscious, one could still argue that the unity is achieved by conscious inference and therefore that the external transfer is not essential to the unity as such.)

To speculate further, if we assume a functionalist view of consciousness, it follows that different systems can have a unified consciousness in virtue of different mechanisms that are apt to support the function of unified consciousness at the temporal scale that is ecologically relevant for the system.

What then is the function of unified consciousness? The answer to this question is notoriously contentious. I argued in 3.1 that the function is to enable the organism navigate the world in a way that is sensitive to both external (states of affairs) and internal context (goals, bodily states). This involves 1) having a perspective - a cognitive structure allowing formation of expectations of how the world will respond to my actions and how my states will change in reaction to a change in the environment; expectations and perceptions that vary with our intentions, 2) being able to combine information from various modular domains in a context-sensitive way.

With this functional description specified, we could finally assess to what extent the leading neuroscientific theories of consciousness explain its unity. The global workspace theory aims specifically at explaining the integration part. At the conceptual level, the unity is explicated as the *unitary* global workspace in which only one message is being broadcast at a time. At the objective level (relations among vehicles of conscious contents), the unity is conceived as a pattern of sustained activation of workspace neurons that connect brain areas over large distances. According to the GW theory, conscious experience is coherent because only vehicles of coherent contents would support each other's activation. Context sensitivity is a result of the mechanism by which a previously unconscious content enters the loop of sustained activation that defines the current global workspace. Given that the modules recruited in the global workspace jointly encode the context, and that a new representation can be recruited to the GW only if it gains activation from it, it is supposedly guaranteed that the changes in the global workspace reflect the changing situation. In short, much of the heavy explanatory work of the GW theory lies in the idea that relevance and coherence is a matter of propensity of a representation to enter the global workspace. To simplify a little bit, this propensity is in turn a matter of variance in strength of neural connections among the candidate representations and the representations already recruited in the current GW.

In explaining why the connections have formed so as to support coherent experience, the global workspace theory refers to the concept of neural darwinism. However, the idea that connections form and get stronger by being involved in

a response causing a beneficial feedback from the environment is too general to provide a satisfactory explanation of relevancy and coherence of conscious contents. The main problem is how to account for incoherent or irrelevant contents. It should not follow from the theory that everything we are conscious of is by definition relevant and coherent, for then the statement would be vacuous. A natural way to allow irrelevant or incoherent contents of consciousness would be to say that they result from facing an anomalous condition - one that leads to a selection of an irrelevant content which nevertheless would be relevant in similar but standard conditions. Such an explanation seems to create more problems than it solves. First, it *prima facie* defers the explanation of the distinction between coherent and incoherent conscious contents to the distinction between standard and anomalous conditions. More importantly however, if the room for irrelevant conscious contents is to be made by appealing to anomalous conditions, how can the GW theory count as explaining the ability to react adaptively to *novel* situations? For novel situations by, in a sense, anomalous.

Despite these challenges to the GW theory, I concluded that it provides a useful framework for thinking about the integration unity of consciousness. Since the GW theory is compatible with both the integration information theory and the predictive coding theory, I proposed to consider the latter theories as complements to the general framework provided by the GW theory. The integration information theory employs the concept of effective information to specify what it means for a group of neurons to underlie an integrated yet differentiated conscious state. Athough using information-theoretic concepts specifies what integration and differentiation mean, it faces a similar problem when it comes to relevance. To justify that the causal relations among states that define the effective information measure are such that the dynamic core always encodes the most relevant information in the current situation, the proponents of the IIT again appeal to neural darwinism. Still, it is not obvious why an assembly or neural groups with the highest effective information should be the one that represents the most relevant contents in the situation, let alone how such representation leads to adaptive behavior. This is where the predictive coding theory offers the most plausible ac-

count thanks to being built on a broader foundation of physical and biological principles.

The PC theory is based on a broader framework that describes principles of self-organisation of organisms using information-theoretic concepts that originated in statistical physics and thermodynamics. The core idea is that the whole organism resists entropy by developing models that enable it to predict sensory consequences of its actions and therefore avoid harmful situations and seek beneficial ones. The predictive coding architecture is a consequence of the organism's need to keep itself in a small number of life-supporting states despite changes in the environment. An important consequence of the PC theory is that both action and perception are explained in terms of the same neural and computational mechanism, for one way of minimizing the prediction error is to act on the world in a way that brings it closer to the system's prediction (the other, of course, is to update one's representation of the world accordingly). Realizing an action is a matter of treating a goal state (conceived ultimately as a set of proprioceptive and interoceptive activations) as given, and performing the bayesian inference of intermediary steps that take us to the goal.[3]

This aspect is particularly important for the unity of consciousness because the seamless connection between action and perception helps us abandon the sandwich model of consciousness - that is the idea that the conscious subject receives inputs as perceptions and issues outputs as actions. In an ecological sense, the subject is nothing less than the whole self-organizing system. In a psychological sense, the subject is the organism's model of itself - its goals, dispositions, characteristics, etc. More specifically for humans, the psychological subject is largely a model of the agent's place in the social world, its social goals, and interactions. Understood this way, the psychological subject is active only in the sense in which the actions are based mainly on the contents that constitute the psychological subject.

[3]Note that the organism seeks to minimize the *long-term* average of surprisal, so the theory accomodates for the short-term increase in prediction error that accompanies the representation of a goal state that is not yet realized, as well as taking actions that lead to short-term discomfort but long-term safety. The proponents of the PC theory introduce precision weighting as the mechanism responsible for the management of prediction errors.

The PC theory does not fully explain the integration unity at the neural level. Rather, the theory states that having a single, coherent, unified, and perspectival view of the world is the goal that the mechanism of prediction error minimization has evolved to achieve. At a time, contents are integrated in the sense that they are congruent with the current hypothesis about the state of the world (including oneself) to the highest possible degree. More precisely, the integration is embedded in the very architecture proposed by the theory: the generative model higher in the hierarchy integrates representations from the lower levels in the sense that it is able to efficiently reproduce patterns of their activation by forming their joint probability distribution. In other words, the integration consists in "recognizing" the probability with which the patterns of activity of lower levels in the hierarchy co-occur.[4] This, however, is an account at the computational level, not material level. Regarding the material level, I argued that the acyclical neural architecture often used for illustrating the neural mechanism that is supposed to realize predictive processing cannot go all the way because that would assume a single predictive node at the top.[5] Without a specific account of how a neural architecture that is not strictly hierarchical realizes predictive processing at high levels of abstraction, the objective account of the integration unity of consciousness is incomplete. The issue is not only technical but conceptual as well: the predictive coding theory cannot count as a satisfactory explanation unless we understand the information flow between different generative models that gives rise to the hypothesis that best explains current precise prediction error (which is the PC theory's reduction of the content of consciousness at a time) and which cannot have the simple acyclical structure. If there is no master generative model, what

[4]As a simplified illustration, consider two lower levels, one representing the number of legs of an animal and the other its skin. These levels can be understood as probability distributions over representations in their domain, i.e. the base-rate probability that an observed animal is quadripedal or furry respectively. A higher-level generative model of animal classification can then be understood as a joint probability distribution sensitive to the fact that the probability distributions of number of legs and skin types are not independent. For example, the joint probability of being furry and having more that eight legs is zero.

[5]Note that even if we hold that the top areas in the predictive processing hierarchy are widely distributed over the cortex, the top node would still be a bottleneck a damage of which should lead to severe decrease of integration.

process organizes the assembly of generative models that jointly explain away current prediction errors? Note the similarity to the problem of how the global workspace is organized over time - how do different, domain-specific modules enter or leave the global workspace.

Despite the conclusion that the PC theory currently does not provide a specific account of this problem, I do not think it is impossible. I outlined a possible answer in 6.2.4: at higher levels of abstraction, the generative models are linked by bidirectional connections (models can form prediction about one another). What might shape the assembly of models that jointly minimize prediction error in a way that achieves the marking flexibility of our thinking is language. The conceptual structure and the coherence constraint embedded in our capacity to speak a language is what allows us to generate novel hypotheses that flexibly explain prediction error in new situations.

From a philosophical point of view, the technical question of what is the neural mechanism and dynamics behind the interaction of generative models is less interesting than the main tenet that cognition and action can be understood as Bayesian inference. I emphasized that the PC theory has not originally been proposed as a theory of consciousness. The inference is an unconscious, computational process using inputs (priors and evidence) to yield an output (posterior probability density). That the brain is an inference machine is ultimately a metaphor - just like the old metaphor of the mind as software running on special hardware that is the brain. It is likely that the metaphor of the mind as a statistician will prove more fruitful for cognitive science than the earlier metaphor of the mind as a computer programme (otherwise there would be little pressure to replace it). Regarding consciousness, however, the metaphor does not bear any obvious advantage.

Throughout this work, I treated the subject and integration unity as two sides of the same coin. Let's turn then to the question how the subject unity emerges from the account summarized above. One of the main implications of the PC theory is that each system or agent will develop generative models that are primarily egocentric - those that will model *his* states, consequences of *his* actions etc. The need for reliable prediction under varied circumstances will motivate de-

velopment of more abstract generative models (e.g. of physical laws), but these will all in the end serve to predict one's own states. Thus a strong side of the PC theory is explaining what Hurley calls perspectival self-consciousness. Experiments on the sense of agency and body-ownership, and their interpretation under the predictive coding framework, discussed in greater detail in 6.2.3, provide a convincing picture of the mechanism underlying implicit bodily self-awareness. The significance of this mechanism is that it is the foundation of self-reference without (conscious) identification. However, to account for the subject unity of consciousness, the picture needs to be extended further from bodily self-awareness to the ability to always reflect on one's contents of consciousness.

I argued that explicit awareness of one's mental states is to be understood as an articulation of the contents of a model specialized in attributing states to oneself (metacognition, self-monitoring, or theory of mind). These models *re-present* some aspects of the current state of consciousness. I proposed that similarly to the way the sense of body-ownership is secured by matching predictions of proprioceptive and interoceptive signals, the sense of identity of the subject of the reflective thought and the lower-order thought is secured by matching predictions of the object state (that is, vehicles of the object state), where the content of the object state is a mental state attributed to oneself and the content of the predictor is its articulation. So, the sense of identity of the subject in the reflective thought 'I am thinking about rules of english grammar.' is a consequence of the fact that the articulated statement correctly matches (that means, generates small prediction error) the content of the lower order thought. To speculate further, if the prediction error is not minimized by the articulation, it could be experienced as struggling to find the right expression of one's content of consciousness; and if the prediction error is high, the subject should not be conscious of any self-reflective thought because the articulation would not get enough support from the object thought to become conscious (it would not be recruited in the global workspace, in terms of the GW theory). Self-attribution of folk-psychological mental states can be understood in the same way as matching predictions made by one of the modules responsible for self-related concepts (e.g. metacognition, theory of mind, or self-monitoring).

It is important to recognize that formation of the models of self-attribution is likely to be mediated by language. Evidence from developmental psychology suggests that the theory of mind develops primarily to predict behavior of other agents and the ability to attribute states to oneself emerges later. The child first learns to read minds of others based on perceptual cues such as facial expression, body posture, or later the situational context. The ability to self-ascribe mental states immediately, without identifying them via the perceptual cues used for other-reading, forms when the child learns to associate the public concepts with interoceptive states based on a feedback provided by competent mind-readers. A concept that was originally applicable thanks to a conscious inference from perceivable cues could thus gain non-inferential and reliable application to oneself - to the point where the unconscious criteria for self-application may be dissociated from the behavioral expression that under standard circumstances warrants the application of the concept to others. That is, the agent may intentionally hide her mental states from others and be perfectly well aware of them.

I also argued that the higher-order reflective states re-present, rather than contain, the lower order states. At the level of vehicles of content (neural activations), there is no qualitative difference between metarepresentation and representation. In computational terms, higher-order states represent some aspect of the lower-order state in a more compressed, or abstract format. The intriguing features of self-reflection, such as self-evidence, transparency, or incorrigibility are thus to large extent products of language, in which the self-reflective content is articulated, and the pragmatics of reporting mental states which establishes the epistemic privilege of the first-person view. The subject that seems to manifest itself in self-reflective thoughts is a concept that owes its existence to the grammar of the language in which we articulate our mental states. It is not an immaterial entity that binds together contents of consciousness and thereby secures their unity. On the contrary, the sense of a unitary subject that permeates our experience is secured by the unity of consciousness that is ontologically prior to it.

This work has been motivated by the idea that the unity of consciousness is the right angle at which the quest for consciousness should be approached. I have tried to describe conceptual obstacles that lie on this path and clear some of them. Reviewing the leading neuroscientific theories showed that much depends on the notion of information integration. The notion has not yet been specified at the neural level at such detail that it could be tested, let alone recreated in an artificial system. The extensive use of information-theoretic concepts in contemporary neuroscience may invite lot of philosophical skepticism based on the critical discussions of classical AI, representational theory of mind, and computationalism. However, the information-related concepts and metaphors that contemporary neuroscience builds on come from statistics and probability theory where 'information' is not be understood as an interpretable symbol. Still, it requires great caution not to slip from informational talk about brain states to ascribing contents to them (I admit I may have slipped few times here). Similarly, it may be extremely difficult to envision how a particular mechanism of information integration described at the neural level could form the unity of consciousness. I suspect that no such account will be completely convincing until it is possible to manipulate the mechanism and consciousness with it, or implement the mechanism in an artificial system and observe a self-reflective agent. And even in that case we might not be any wiser regarding the question of what is the foundation of consciousness. It may well happen that we will know what process is responsible for consciousness without getting any deep understanding from it.

Bibliography

Apps, M. A. and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews*, 41, 85–97.

Arviv, O., Goldstein, A., and Shriki, O. (2015). Near-critical dynamics in stimulus-evoked activity of the human brain and its relation to spontaneous resting-state activity. *The Journal of Neuroscience*, 35(41), 13927–13942.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bayne, T. (2010). *The Unity of Consciousness*. Oxford ;Oxford University Press.

Bayne, T. J. and Chalmers, D. J. (2003). What is the unity of consciousness? In A. Cleeremans (Ed.), *The Unity of Consciousness*. Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 18(2), 227–247.

Brook, A. (1997). *Kant and the Mind*. Cambridge University Press.

Brook, A. and Raymont, P. (2017). The unity of consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Carr, C. and Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *The Journal of Neuroscience*, 10(10), 3227–3246.

Carruthers, P. (1998). *Language, thought and consciousness: An essay in philosophical psychology*. Cambridge Univ Press.

Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.

Castaneda, H.-N. (1966). 'he': A study in the logic of self-consciousness. *Ratio*, 8(December), 130–57.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3), 200–219.

Chalmers, D. J. (1997). Availability: The cognitive basis of experience? In N. Block, O. J. Flanagan, and G. Guzeldere (Eds.), *The Nature of Consciousness* (pp. 148–149). MIT Press.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, 23(2), 149–178.

Clark, A. (2001). *Mindware*. Oxford University Press.

Clark, A. (2009). Spreading the joy? why the machinery of consciousness is (probably) still in the head. *Mind*, 118(472), 963–993.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204.

Damasio, A. (2012). *Self comes to mind: Constructing the conscious brain*. Vintage.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1), 1–37.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1991). *Consciousness Explained*. Penguin.

Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

Diaz, R. and Berk, L. (1992). *Private speech: from social interaction to self-regulation*. L. Erlbaum.

Dutton, D. G. and Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of personality and social psychology*, 30(4), 510.

Edelman, G. M. (2003). Naturalizing consciousness: a theoretical framework. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9), 5520–4.

Flavell, J. H., Green, F. L., and Flavell, E. R. (2000). Development of children's awareness of their own thoughts. *Journal of Cognition and Development*, 1(1), 97–112.

Fletcher, P. C. and Frith, C. D. (2009). Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48–58.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Bradford Bookks. A BRADFORD BOOK.

Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard University Press.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revisited: The Language of Thought Revisited*. OUP Oxford.

Francken, J. C. and Slors, M. (2014). From commonsense to science, and back: The use of cognitive concepts in neuroscience. *Consciousness and cognition*, 29, 248–258.

Franklin, S. (2003). A conscious artifact? *Journal of Consciousness Studies*, 10(4-5), 47–66.

Franklin, S. and Graesser, A. (1999). A software agent model of consciousness. *Consciousness and cognition*, 8(3), 285–301.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, 9(10), 474–480.

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual review of neuroscience*, 32, 209–24.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gazzaniga, M. (1985). *The social brain: discovering the networks of the mind*. Basic Books.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: does the corpus callosum enable the human condition? *Brain : a journal of neurology*, 123 (Pt 7, 1293–326.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193(2), 559–582.

Gopnik, A. (2009). *The philosophical baby: What children's minds tell us about truth, love & the meaning of life*. Random House.

Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, 4, 383–427.

Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56(5), 872.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3.

Hurley, S. L. (1998). *Consciousness in Action*. Harvard University Press.

Hurley, S. L. (2003). Action, the unity of consciousness, and vehicle externalism. In A. Cleeremans (Ed.), *The Unity of Consciousness* (pp. 78–91). Oxford University Press.

Husserl, E. (2013). *Zur Phänomenologie des inneren Zeitbewusstseins: mit den Texten aus der Erstausgabe und dem Nachlass*, volume 649. Meiner Verlag.

Irvine, E. (2012). *Consciousness as a scientific concept: a philosophy of science perspective*, volume 5. Springer Science & Business Media.

Kant, I., Guyer, P., and Wood, A. (1998). *Critique of Pure Reason*. Oeuvre. Cambridge University Press.

Kant, I., Smith, N., Caygill, H., Banham, G., and Smith, N. (2016). *Critique of Pure Reason, Second Edition*. Palgrave Macmillan UK.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Keller, P. (2001). *Kant and the Demands of Self-consciousness*. Cambridge University Press.

Kitcher, P. (1993). *Kant's transcendental psychology*. Oxford university press.

Lillard, a. (1998). Ethnopsychologies: cultural variations in theories of mind. *Psychological bulletin*, 123(1), 3–32.

Lupyan, G. and Clark, A. (2015). Words and the world predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, 24(4), 279–284.

Ma, W. J., Körding, K., and Goldreich, D. (2013). Bayesian modeling of perception.

Nagel, T. (1971). Brain bisection and the unity of consciousness. *Synthese*, 22(3), 396–413.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641–1646.

Penny, W. (2012). Bayesian models of brain and behaviour. *ISRN Biomathematics*, 2012.

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, 134, 17–35.

Price, C. J. and Friston, K. J. (1997). Cognitive conjunction: a new approach to brain activation experiments. *Neuroimage*, 5(4), 261–270.

Prinz, J. (2006). Is the mind really modular. *Contemporary debates in cognitive science, ed. RJ Stainton*, (pp. 22–36).

Proust, J. (2003). Does metacognition necessarily involve metarepresentation? *Behavioral and Brain Sciences*, 26(03), 352–352.

Ramsey, W., Stich, S., and Garan, J. (1990). Connectionism, eliminativism, and the future of folk psychology. In *Philosophy, Mind, and Cognitive Inquiry* (pp. 117–144). Springer.

Ransom, M., Fazelpour, S., and Mole, C. (2017). Attention in the predictive mind. *Consciousness and cognition*, 47, 99–112.

Rosenberg, J. (2005). *Accessing Kant: A Relaxed Introduction to the Critique of Pure Reason*. OUP Oxford.

Schooler, J. W. (2002). Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6(8), 339 – 344.

Schwitzgebel, E. (2011). *Perplexities of consciousness*. Life and Mind : Philosophical Issues in Biology and Psychology Series. MIT Press.

Schwitzgebel, E., C., H., and Y., Z. (2006). Do we dream in color? cultural variations and skepticism. *Dreaming*, 16, 36–42.

Searle, J. R. (1990). Is the brains mind a computer program. *Scientific American*, 262(1), 26–31.

Sellars, W. S. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, 1, 253–329.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11), 565–573.

Seth, A. K. (2015). The cybernetic bayesian brain. In T. K. Metzinger and J. M. Windt (Eds.), *Open MIND* chapter 35(T). Frankfurt am Main: MIND Group.

Shanahan, M. (2005). Global access, embodiment and the conscious subject. *Journal of Consciousness Studies*, 12(12), 46–66.

Shew, W. L. and Plenz, D. (2013). The functional benefits of criticality in the cortex. *The neuroscientist*, 19(1), 88–100.

Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65(October), 555–67.

Siegler, R. S., DeLoache, J. S., and Eisenberg, N. (2011). *How Children Develop*. Macmillan.

Simons, D. J. and Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9), 1059–1074.

Smolensky, P. (1995). On the projectable predicates of connectionist psychology: A case for belief. In C. Macdonald and G. F. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation*. Blackwell.

Sperry, R. W. (1968). Hemisphere deconnection and unity in conscious awareness. *American Psychologist*, 23(10), 723.

Strawson, P. F. (1966). *The Bounds of Sense: An Essays on Kant's Critique of Pure Reason*. Methuen.

Suzuki, K., Garfinkel, S. N., Critchley, H. D., and Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, 51(13), 2909–2917.

Taylor, J. (2012). The problem of'i':: A new approach. *Journal of Consciousness Studies*, 19(11-12), 233–264.

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, 5, 42.

Tononi, G. and Edelman, G. M. (1998). Consciousness and Complexity. *Science*, 282(5395), 1846–1851.

Wegner, D. M. (2005). Who is the controller of controlled processes? In R. R. Hassin, J. S. Uleman, and J. A. Bargh (Eds.), *The New Unconscious. Oxford Series in Social Cognition and Social Neuroscience* (pp. 19–36). Oxford University Press.

Wittgenstein, L. (1958). *The Blue and Brown Books*. Harper and Row.

A Fallibility of Introspection

In this section, I will recount few empirical studies and related arguments which lead to the conclusion that introspection is fallible and theory-laden. Introspection is here understood in its limited sense of forming a higher-order mental state that is about the content of consciousness. Saying that introspection is theory-laden means that the content reflected on in the higher-order state is not somehow given, unchanged from the way it is presented in the lower-order mental state. Rather, the higher-order represented content is a conceptualization of the lower-order content.

Sellars (1956) famously provided a thorough argument against the myth of the given, and one of its conclusion is that introspection (introspection, as the target of his argument, could be understood as the purportedly direct acquaintance with sense-data that Russell built on) does not present us with an immediately given content, for nothing can both have un unassailable epistemic warrant usable in justifying other propositions and be epistemically independent (i.e. fundamental) of the propositions to which it is supposed to be inferentially related. His argument is a priori - it shows that the conception of introspection as direct, conceptually unmediated awareness of one's conscious contents cannot be right, especially insofar it is supposed to play the role of epistemic foundation. The a priori argument, however, does not show how exactly can introspection be wrong and what could be the criteria for judging whether an introspective report is right or wrong. Discussion of specific cases may therefore give us better understanding of how introspection can be wrong.

A.1 Do we dream in color?

Proponents of infallible introspection often give examples of introspecting perception of some basic visual features, e.g. colors or shapes. Since vision is the dominant sensory modality in our experience, it seems very unlikely that we could be confused as to whether some visual experience is in color or not - especially on the assumption that we have an immediate (i.e. non-inferential, non-cognitive) access to this visual quality. But this confusion actually arises with dreams. Schwitzgebel (2011) found out that people began to report dreaming in color when color televisions first appeared in average american households in the early 1960s. So, while in the era of monochromatic TVs only 9-29% of people reported dreaming in color, after color TV was introduced the ratio increased to 81-100%.[1] This can hardly be a coincidence. Nevertheless, it could be argued that this does not show that the introspective judgement about dreaming in color is fallible, for it could well be the case that people *really* started dreaming in color when they got exposed to color TV. How can we tell whether the change in reports is an effect of different interpretation or an effect of different experience caused by the increased exposure? Note that if we assume that changing our conceptual framework through which we introspect *ipso facto* changes our experience, then these alternatives coalesce - they make no difference. So, looking at arguments that disambiguate between the two alternatives will help us understand what it takes to correct introspection.

Let's first argue against the view that dreams really changed from black and white to color. It is not obvious why increased exposure to color media should change the visual quality of dream experience from monochromatic to color given that people's everyday visual experience was in color anyway. We would have to assume a very special relation of TV to dream experience to account for this change. Without a compelling reason, it is more plausible to assume that dream experience is based on everyday perceptual experience, not a specific subset of experience (media watching) that, in addition, we have been having only recently, compared to dreams.

[1] See(Schwitzgebel, 2011, chpt. 1) for a review of studies of dreams conducted between 1933 and 2008.

This negative point is supported by the result of a follow-up study that specifically looked at the correlation between exposure to color media and color dream reports. Schwitzgebel et al. (2006) studied different socio-economic groups in China which differed in their access to technology and hence in exposure to color media. They found that while *individual* exposure was only weakly correlated with color dream reports, the correlation was stronger at the group level. Accordingly, Schwitzgebel draws the following conclusion:

> These results suggest that whatever is affecting people's reports is something shared at the group level - something, I suspect, like cultural attitude, or the availability of certain metaphors, or certain ways of thinking and talking about one's dream life. (Schwitzgebel, 2011, p. 7)

So, unless we bite the bullet and say that dream experience is *really* influenced by TV and socio-economic status, we have reasons to doubt that people know whether they dream in color or not.

Now, it could be argued that this does not mean that introspection is fallible, for there is nothing to be known - there is no fact of the matter whether the participants dreamt in color or black and white besides what they say. The reason why the introspective judgement is false is not because the person *really* dreamt in color while reporting dreaming in black and white (or vice versa). It is false because the dream experience may lack the color dimension altogether and hence be indeterminate in this respect. The introspective judgement is false because its presupposition (things seen in dreams must have determinate color) is false. The situation is, I think, similar to the indeterminacy in literary fiction. The works of A. Conan Doyle probably do not mention the color of Sherlock Holmes's socks, hence it is indeterminate. If a reader is asked what color his socks are and if she conceptualizes Holmes as a real character, she will be inclined to give a determinate color, probably one that befits the character. Similarly, if we conceptualize the dream experience as a seen film, for example, it will come natural to us to fill in the missing color dimension.[2]

[2]What it is about dreams that invites the film metaphor? A speculative answer would be that it is the very passive character of dreams that makes us conceptualize them this way. Except

Finally, could we decide whether people dream in color or not independently of their reports? There seems to be a catch: independent means of deciding whether some experience is in color or not (e.g. by functional imaging of V4 or other cortical area) would ultimately rely on subjective reports anyway, for to establish that some objective finding (a neural correlate of consciousness) is a reliable indicator of color experience, one has to take into account subjective reports.[3] So in order to falsify introspective judgements by objective means, we need to rely on them first. But this is no more of a paradox than that in order to say that someone is lying we need to assume he is capable of telling the truth. First-person mental ascriptions have strong but not unquestionable authority.

A.2 Inconsistency between introspective reports and folk-psychological beliefs

Another, more general, finding which puts the infallibility of introspection in doubt is the fact that people differ considerably in their reports of the quality of experience but their performance in the respective area is quite similar. Schwitzgebel (2011) reviews numerous studies of visual imagery in an attempt to shed some light on the question to what extent is thought imageless. He found that while people differed in their subjective reports of the visual richness of their imagery and thinking, their performance in tasks like mental rotation, or visual creativity was only weakly or not at all correlated with the reported richness. This

for the rare cases of lucid dreaming, the subject has no feeling of control over the contents of the dream - it unfolds before our mind's eye. This gets some support from Hobson's AIM (Activation, Input source, Modulation) model of states of consciousness which characterizes REM stage (where dreams occur) as being low in modulation dimension, meaning that the experience is that of a passive observer.

[3]It could be argued that cognitive neuroscience can do without introspective reports by making assumptions about the subject's conscious contents independently of their introspective reports, e.g. by manipulating their attention and assuming that what they attend to, and differentially react upon, must be conscious. But even in that case the interpretation would rely on introspective reports that are indirectly related to the hypothesized contents. For example, the experimenter needs to ascertain that the subject understands the task, that she is willing to comply with it (and not ruin the experiment by giving haphazard answers), etc.

is a striking on the assumptions that 1) our conscious performance in a task is guided by our conscious contents, and 2) introspection is infallible (hence people reporting visually rich imagery during the task really have more visual experience). In theory, we could argue against 1), saying that one's phenomenology is independent of the representations that actually underlie the cognitive tasks. But this would render phenomenology epiphenomenal and explanatorily inert. Thus a more plausible account for the inconsistency seems to be saying that people differ largely in the way they conceptualize their thoughts and imagery, and not so much in the character of the representations which underlie the cognitive task and which the participants try to describe. For some people it is unthinkable that one could have a purely abstract thought, not accompanied by any image, while for others it is perfectly conceivable. As Schwitzgebel points out, the debate goes at least as far as the controversy about abstract ideas between Locke and Berkeley, who disagreed on whether one could entertain an idea of a triangle that has no specific shape.

The idea that introspective reports depend substantially on one's conceptualization of experience is further corroborated by another experiment which Schwitzgebel designed to assess the controversy between so-called abundant and sparse views of consciousness. The sparse view holds that we are conscious of only a few things that we attend to at a time, whereas the abundant view holds that we are conscious of a wide perceptual and emotional field even outside the focus of our attent. The controversy should be easily resolved if introspection provided us with unproblematic access to our conscious experience: we could just look inside and tell reliably (and consistently over time) whether we are conscious of many diverse things or not. The heat of the debate between these two views indicates that people's opinions differ considerably in this respect. And it is more plausible to assume that people differ in the way they conceptualize their experience rather than in their consciousness, especially when there is little difference in performance.

In the actual experiment subjects wore beepers during the day and were asked to describe the content of consciousness at the moment when the beeper went off. This setup let people go on in their everyday activity, thus allowing for sampling

the content of consciousness as it unfolded naturally, outside the experimental room where consciousness is often contaminated by attending to the demand characteristics. After each day, Schwitzgebel interviewed the subjects to find out more about the situation they were in when the beeper went of, so that he could infer what the subject was paying attention to. Interestingly, he also discussed the sparse/abundant view controversy with the participants and asked them to articulate their position in the debate before and after the experiment.

It turned out that participants' reports were not entirely consistent with their view (sparse/abundant) - every participant reported some experience in an unattended modality (against the sparse view) and at the same time did not report experience in some unattended modality that was nevertheless stimulated, e.g. tactile experience in one's foot (against the abundant view). Consequently, after the experiment subjects shifted their position toward a middle-of-the-road view. Such a moderate view, however,is difficult to accomodate by current psychological theories.[4] This is a rather baffling result. On the one hand it might seem that subjects learned from their reports and adjusted their theoretical view (from one extreme - sparse or abundant - to a moderate view), thus suggesting that our folk-psychological view consciousness is indeed informed by introspection. On the other hand, the prior opinion about the richness of conscious experience correlated only weakly with the reported richness. The small effect could be a result of a general motivation to be consistent with the previously professed view of consciousness: if I commit explicitly to the sparse view of consciousness, I might deliberately omit to report some marginally conscious contents just to comply with the view. I would argue that the shift in the theoretical view was due to explicit theoretical discussions with the experimenter in which he inevitably provided the participants with new ways to think about their experience.[5]

[4]The sparse view draws naturally on theories of attention that have been developed without any explicit background theory of consciousness. Likewise, the abundant view can be seen as identification of consciousness with supraliminal perception. A moderate view does not find any such clearly defined psychological concept to be based on (safe for an interesting but quite speculative concept of diffuse attention). For an interesting discussion of this topic, see (Schwitzgebel, 2011, chpt. 5).

[5]Consider the following step in the experiment taken by Schwitzgebel. In order to disambiguate whether by reported absence of visual experience the subject meant "real" absence or

A.3 Introspective children and cultural differences

Schwitzgebel presents other examples of putative failures of introspection. The previous discussion suffices to show that introspection is fallible and what kind of arguments can demonstrate it. Let me conclude with two more general findings.

First, an experiment by Flavell et al. (2000) suggests that the ability to introspect develops relatively late, after the child learns the theory of mind and masters concepts of various mental states. Children of age 5 and 8 (and adult controls) were asked not to think about anything for 30 seconds. In a follow-up interview, most 5-year-olds denied having any thoughts or mental activity, while most 8-year-olds (and all adults) "correctly" admitted having some thoughts. The authors carefully ruled out the possibility that the denial may be an effect of the children's tendency to comply with experimenter's order. They also rule the possibility that it is an effect of failed recall from, or encoding, in memory. The conclusion they draw is that 5-year-olds did "have potentially noticeable, conscious thoughts but they were less able or less disposed than the older participants to notice them."[6] When speculating about what makes 8-year-olds better in this respect, the authors point to formal schooling system which, starting with elementary school, leads children to pay attention to their mental activities in problem solving. Although this experiment is neutral in respect to the question whether introspection yields some immediate knowledge of internal states or whether it is theory-laden, it suggests that introspection is an unusual and possibly culturally driven practice.

This idea gets further support from ethnographic studies that show cultural differences in introspection and the theory of mind reasoning. As Lillard (1998) observes, there are cultures, such as Kaluli in Papua New Guinea, which regard

experience of blackness, he introduced the concept of "phenomenal blindness" and when this was understood, he asked the question "Could a phenomenally blind person, a twin of you in all respects except lacking visual experience, have had the same conscious experience at that moment?" (Schwitzgebel, 2011, p. 101) No doubt, just understanding this question requires a set of conceptual distinctions that most people would probably never come up with spontaneously.

[6](Flavell et al., 2000, p. 108)

minds as unknowable and do not attempt to ascribe mental state to others. This is reflected in their moral judgement that is based solely on consequences of an act, not the intention behind it (which seems to be typical for our culture). This fact renders it less likely that the ethnographer's claim about the lack of mentalizing activity is a crude misintepretation.

A.4 Conclusion

Previous discussion showed concrete examples of how introspection can fail and how the way we conceptualize experience may influence our reports of conscious contents. This conclusion is important for the discussion of the unity of consciousness in that it undermines an otherwise attractive idea of the mind being transparent to itself. Besides giving empirical support for Kant's argument that we know our mind via the inner sense only as it appears to us, not as it is of itself, the previous discussion also showed what kind of reasons can justify saying that someone's introspective judgement is false.

It follows that the higher-order introspective state is different from the lower-order object state: it is not the case that the latter is a proper part of the former, as would be the case, for example, if we conceived of introspection on the model of a specific propositional attitude attached to some content (e.g. 'I think x' where x ranges over introspectible states). As a consequence, we need to account for the sense of ownership of our thoughts (the subject unity of consciousness) in terms of relations among higher-order states and object states, not in terms of the content of higher-order states. This, I think, is yet another variation on the Kantian idea that the unity of consciousness is not consciousness of the unity.

B Measuring information integration

Tononi (2004) suggests that consciousness corresponds to the capacity to integrate information and offers the following information-theoretic measure of integration. Let's have a system S composed of n units carrying information. Let A, B be subsets of S. Next, we introduce the measure of effective information EI between A and B, which is supposed to capture the extent to which these subsets influence each other's informational states. EI is defined directionally: the effective information of A on B is mutual information MI between all possible states of A and the states of B that arise as a result of those states in A. Mutual information is a common informational-theoretic measure defined as MI(A,B) = H(A) + H(B) - H(AB), where H(A) is information entropy, defined in turn as $H = -\sum p_i * \log p_i$ where p_i is the probability of the i-th state of A. The functional form for entropy implies that entropy is maximimized for uniform probability distribution, i.e. one in which each state occurs with equal probability. Such a system is the least predictable (most uncertain), hence most entropic or disordered. Effective information from A to B can then be formulated as MI(A^{max}, B), where A^{max} corresponds to states of A with maximal entropy (which is just another way of saying that we take every possible state of A and see what state in B arises). Finally, the non-directional definition of effective information is just the sum of effective information from A to B and vice versa: $EI(A \rightleftharpoons B) = EI(A \rightharpoonup B) + EI(B \rightharpoonup A) = MI(A^{max}, B) + MI(B^{max}, A)$.

Now, the capacity of S to integrate information is defined as effective information between such bipartition of S into subsets A,B that has the lowest normalized effective information. To see why, consider the case in which S can be decomposed into two informationally (causally) independent subsets. Since there is no

information transfer between them ($EI(A \rightleftharpoons B) = 0$), they are independent systems that do not integrate any information. Recall again the example with a camera chip of 1000 photodiodes. Although the chip as a whole can discriminate 2^{1000} states, the information is no more integrated than in case of 1000 human observers around the world, each reporting on a local state without any interaction with other observers. Similarly, retinal neurons or early visual areas with local, independent receptive fields integrate little information. Tononi argues that this is the reason why low-level feature processing in the brain is unconscious. Finally, EI needs to be normalized by the minimal information entropy available to A or B, otherwise the bipartition with minimal EI would be one in which A or B is composed of just one unit.

The capacity for information integration of S is thus defined as effective information across its weakest link. Finally, the theory holds that the content of a conscious state corresponds to the activity in the subset S of the whole neural system that has the highest amount of integrated information. This subset is called the dynamic core. Tononi (2004) uses examples of simple networks to show that it is unlikely that the whole network would be this subset with maximal integrated information. Intuitively, including specialized subsystems that process information relatively independently and only pass their result to the rest of the network would result in a decrease of effective information. In the discussion of the global workspace theory, we noted that the informal criterion of recruiting a neural group into the global workspace is the relevance of its representation. Integrated information theory can be interpreted as showing that relevance can be specified as effective information of that neural group and the rest of the system.[1] If we look at the brain as an isolated system (as the information integration theory does), then effective information among neural groups may seem to capture very little of the sense of ecological relevance used in the GW theory. However,

[1]More precisely, the reason for "recruiting" a neural group into the global workspace would not be the effective information between it and the current GW but rather the fact that union of the neural group and the subsystem that currently forms the GW will attain a higher measure information integration than the subsystem alone. However, this situation can occur only if the effective information between that neural group and the current GW is high, relative to other candidates competing for access to the GW.

since informational (causal) relations among neural groups are formed over time by interacting with the environment and various neural learning mechanisms, we can say that parts of the neural system have informational relations they have *because* they proved to be ecologically relevant.