Creative introspection and the structure of the experiencing subject

Silvère Gangloff

September 22, 2021

> *"Another observation. It never happens that the unconscious work gives us the result of a somewhat long calculation all made, where we have only to apply fixed rules. [...] The rules of these calculations are strict and complicated. They require discipline, attention, will, and therefore consciousness. In the subliminal self, on the contrary, reigns what I should call liberty, if we might give this name to the simple absence of discipline and to the disorder born of chance. Only, this disorder itself permits unexpected combinations."* - **H.Poincaré**.

I.– Statement

The word '*consciousness*' may refer to multiple realities: while some theoricians think of it as the fact for a physical system to have phenomenal experience and search to explain why certain physical systems have experience and others do not, others focus more on the relation of the subject of phenomenal experience to this experience, and explain consciousness by what happens in the brain when some information becomes conscious. Others think of consciousness in terms of attribution of phenomenal experience to another physical system in the subject's own experience. The multiple aspects of consciousness differ in the possibility and difficulty of analysis, however the easiest way is not necessarily the most meaningful. My point of view on this is that thinking about consciousness through some reading grid which allows the progression from a description of simple but fundamental aspects of it to a complexification into a more detailed description and analysis might end with unexpected insight into other aspects of consciousness that are more difficult to approach and a priori out of reach. This is why when thinking about consciousness I choose to focus on what I shall call the *experiencing subject*, meaning the subject in its activity on and in its phenomenal experience, in which it *navigates*, that it *conceptualises*, from which it *extracts* information, and uses it to *transform* this experience (a priori this list is not exhaustive but should give an idea of what I am talking about). From this point of view, one may conceive other aspects consciousness such as that the degree of consciousness of a subject which should be related to the 'richness' of the *conceptualisation* of experience (constructed out of the execution of a certain series of transcendental operations). Describing the *structure* of the possible dynamics of the experiencing subject, in other words the form of the subject's relation to its experience - which as such participates to the concept of consciousness as such, as it is not specific to any experience in particular - may allow a correspondance of this structure with the structure of the humain brain, which should be expected to be partial. Furthermore in order for this correspondance to be possible, the description should come with a conceptualisation of this structure, an understanding of it.

The purpose of this article is not to present a theory of consciousness for which the author could enter the competition for attention and fundings; instead it appears to me that the right way to conceptualise consciousness is not clear yet and that it is more important for understanding it to explore methods of conceptualising than producing statements about it, of which a theory is expected to be made of. For the definition of the general method, I found that B.Josephson and B.Rubik, in the a report on the *1992 Athens Symposium on Science and Consciousness* [JR92], present accurately the difficulties of a science of consciousness, in particular when submitting oneself to 'constraints on thinking about consciousness that might be imposed by conventional modes of thought'. They arrived in particular at the two following conclusions:

"*We need actively to address the limitations of scientific approaches, verification, and theories, and to find a place in our world view for personal knowledge gained through introspection.*" **[1]**

"*Language itself can provide an effective means of exploring quasi-objectively what has previously been characterised as being purely subjective.*" **[2]**

Although I was not aware of this report before writing down the present text, the other article I wrote last year, titled *A formal window on phenomenal objectness* [G20], was partially devoted to conclusion **[2]**, in particular the question of how to construct a collective discourse about consciousness and how it is related to the activity of the (human) brain, that would consist in reliable knowledge. I formulated in this article the idea that this construction should involve a dialogue between the disciplines of mathematics [for to connect the structure of the experiencing subject and the structure of the brain, one has to use mathematical language] and of philosophy [as consciousness is a typical object of philosophy, with corresponding difficulties]. Along a history of demarcation, these disciplines acquired a certain form which make them practically incompatible, and the dialogue between them difficult. In order to re-form this dialogue I analysed the co-definition of mathematics and philosophy in relation with how they make use of language, that I interpreted using some notions of *statical* and *dynamical* designations that I introduced. Abstracting the factual and historical separation of mathematics and philosophy in terms of the compartimented use of these two types of designations, one may see that mathematics and philosophy can take forms that can coexist in the same (meaningful) discourse. I was then able to build on this conceptualisation a way to approach one fundamental aspect of consciousness identified in *Integrated information theory* and that I chose to rename *phenomenal objectness.*

Although there is certainly a lot more to say about the use of language to talk about consciousness, my point in the present text is to address the conclusion **[1]**. The status of introspection seems to root a critical division between theoretists and experimentalists. As S.Dehaene expresses it [D13], the period of behaviorism in cognitive sciences and its rejection of introspection and consciousness as an object of study left a mark, and it is only recently that consciousness has been reconsidered as a serious object of study in cognitive sciences. However experimentalists - S.Dehaene in particular - reconsidering introspection often take it as an object of inquiry, in restricted situations in which it can be properly defined, and not as a method for this inquiry. In fact this kind of restriction may come with a restriction of expressive power. When taken as a method of inquiry, introspection is often only considered, when properly used, as a way to collect necessary data (see for instance [BP13]). What experimentalists often forget is that the field of mathematics provides a clear example of construction of clear concepts and reliable knowledge on them - meaning not only data - coming out of the use of introspection: in a sense one can say of mathematics that they are a particular form of '*creative introspection*' - creative of concepts.

I think that the main critic addressed to introspection as a method of inquiry is that it distorts its object, leading in particular to false judgements on it. In fact I find this critic incoherent for the reason that any inquiry of the reality, wether theoretical or experimental, involves a distortion of what is taken as object of study - the object as it is perceived - replacing it in the mind with another set of objects and relations. Also it often happens that even careful experimental inquiry arrives at false judgements about its object (due to manipulation or interpretation errors) - however these errors are only temporary as long as one is able to look into the protocol and its realisation and interpretation in order to correct the errors. In fact this distortion of the object does not really matter, as long as ultimately the process of inquiry leads to the observable augmentation of the battery of (meaningful) concepts available to analyse the reality.

My point of view is that an understanding of consciousness can only come with its conceptualisation, and for that the use of introspection is necessary. The real question is then not *shall we use introspection* ? but how to use it ? In answering this question it might be of interest to look at how introspection is used in the field of mathematics, what are the conditions (in particular self-constraints) that lead this use of introspection to meaningful concepts, and how to generalise this use to other objects of introspection, such as the structure of the experiencing subject. This generalisation does not have to involve at least in the beginning - and in fact one should not expect it - the importation of language structures present in the field of mathematics as it is (for instance, axioms, definitions, theorems and their proofs, or even principled theorems), such as in *Integrated information theory* or *Predictive coding theory* for instance.

As a matter of fact the practice of mathematics deals closely with an aspect of introspection which makes its use difficult, which is the chaoticity of what appears in the mind along with a particular *designation*. What I call designation here (in the same way as in [G20]) is the action of making present a certain 'area' of cognition, which happens when we pronouce a word for instance. The word '*consciousness*' is an example of *dynamical designation*, meaning that with it many things may be present in my mind when I think about consciousness or discuss about the nature of consciousness with another person, and it is not clear how they are structured, what amongst these seem to characterise what it is, etc; moreover they appear chaotically, without a rational order. This makes it difficult in particular to introspect about this designation, but not impossible in principle (in fact it is by introspection that experimentalists such as S.Dehaene can come up with a clean definition of a particular type of introspection, for instance conscious access to an information). This is the chaoticity that H.Poincaré, in the quote above, locates in the '*subliminal self*', which I believe corresponds to what philosophers since Plato have called χώρα, initially (before Plato) designating the territory of the πόλις outside of the city, an intermediate between its inside and outside. Plato mapped the structure of the πόλις to what I called the structure of the experiencing subject, designating by the term χώρα what lies between the intelligible and the sensible. My interpretation of χώρα is that it is the cognitive 'place' where mental intelligible content is formed and unformed, in particular by integrating or disintegrating partially conscious elements; in line with H.Poincaré, this is the place where conceptual creation is possible (in particular in mathematics). The difficulty of dwelling in the χώρα is that it does not appear in the same way as objects (the type of beings we are more used to deal with conceptually) appear, or any experience: in fact is a place of particular type of experiences, where objects disappear and appear; it seems as chaos from the point of view of the conscious center, the πόλις, but appears progressively ordered in its own way (as deterministic chaos tells us) when dwelling in it. In fact I think that χώρα should be the object, the place as well as the method of any theoretical study of consciousness, because that is where the dynamics between the conscious and the unconscious happen, thus delineating the conscious - as well as the unconscious. Any theory which does not is not actually talking about consciousness as such but something else. As a matter of fact the chaoticity of χώρα enters in contradiction with a method of inquiry defined by pre-determined and well-defined rules. On the other hand, although mathematicians may differ on the degree to which they accept how much unconscious is involved in reasoning, I believe that any mathematician would agree with H.Poincaré that there is a part of unconscious work in mathematical creation. The way H.Poincaré describes this creation is related to the way mathematicians find the solution to a problem, observing that the process of finding a solution consists first in a conscious 'ingestion' of the elements of a problem, and some failed attempts to search for a solution which extends directly the methods developped to solve other problems. Then the work is the one of the unconscious at the end of which the solution 'appears' with extreme clarity, with only left the conscious work to write it down in a structured way, and verify it by checking carefully every aspect of it. This is the kind of illumination that mathematicians often like for it comes with a certain sense of beauty, but the apparition of a solution can appear more progressively.

However chaoticity appears only in the individual introspection and not at the level of the collective discourse, which is meticulously bound to consist in the presentation of well-formed solutions to problems and definitions of concepts which appear useful along with the search for solution. What may repell experimentalists from the use of introspection in cognitive science is that it is often rather shallow and short before results of introspection are presented. Contrarily to common belief, I think that the early exchange of ideas, before they are carefully and deeply weighted, can only create confusion in the mind and the one of the other; furthermore it seems that the rush into answering to the critics act on the development of the conceptualisation itself, being determined by the dialogue and the psychology of theoricians more than the reality, their need to be right over the other, holding certain ideas not because of their expressive power but because they answer the critics, in particular when these critics come from a non-understanding of the initial presentation of the theory. Concepts formed are thus not describing adequately this reality but are conceived only to construct a presentable theory, which may become more and more obscure. In fact it may also actually be clear that a discourse which is substantially formed as a response to the other can not be closer to the truth than this other. Experimentalists may see in the method of inquiry, introspection thought as the shallow and straightforward judgement after 'looking inside', the origin of this confusion, but I think this is due mainly to the way introspection is executed and how early the dialogue is involved in the constitution of a conceptualisation. When it is not directly involved, the theorician is able to put his own ideas in competition, thus detaching himself or herself from each of them, and the failure of an idea is not the failure of the theorician or his or her way of thinking. Chaoticity is delt with by the individual theorician and does not appear in the collective discourse. Maybe this type introspection should be differenciated from straightforward introspection - maybe one could name it introrelation. In the remainder of this text however I will stick to the term introspection, for I believe that this is the right use of the word.

On the contrary it seems that chaoticity appears directly in the discourse about consciousness, and this problem has been adressed by B.Josephson and B.Rubik. To illustrate the problem, I would like to quote a schematic dialogue written by D.Chalmers [C18] between realists and illusionists about consciousness:

"***Realist:*** *People obviously feel pain, so illusionism is false.*
Illusionist: *You are begging the question against me, since I deny that people feel pain.*
Realist: *I am not begging the question. It is antecedently obvious that people feel pain, and the claim has support that does not depend on assuming any philosophical conclusions. In fact this claim is more obvious than any philosophical view, including those views that motivate illusionism.*
Illusionist: *I agree that it is obvious that people feel pain, but obvious claims can be false, and this is one of them. In fact, my illusionist view predicts that people will find it obvious that they feel pain, even though they do not.*
Realist: *I agree that illusionism predicts this. Nevertheless, the datum here is not that I find*

it obvious that people feel pain. The datum is that people feel pain. Your view denies this datum, so it is false.
Illusionist: *My view predicts that you will find my view unbelievable, so your denial simply confirms my view rather than opposing it.*
Realist: *I agree that my denial is not evidence against your view. The evidence against your view is that people feel pain.*
Illusionist: *I don't think that is genuine evidence.*
Realist: *If you were right, being me would be nothing like this. But it is something like this.*
Illusionist: *No. If 'this' is how being you seems to be, then in fact being you is nothing like this. If 'this' is how being you actually is, then being you is just like this, but it is unlike how being you seems to be."*

Here is the way I see it:

Realist: *I think you are wrong.*
Illusionist: *Hmmm. No I don't think so.*
Realist: *I aknowledge what you say, but I definitely think you are wrong.*
Illusionist: *God damnit, I am right.*

and so on.. In other words, a dialogue between theoricians who, relying only on straightforward introspection to respond to critics without a prepared discourse that is the fruit of retroaction on his or her own discourse - in particular examining the coherence of the ideas, the significance of the concepts used, their relations, the structuration of the discourse around few central concepts, etc. - has more chance to be reduced ultimately, along with the dialogue, and despite the apparent diversity of the arguments, in a simple repeated affirmation. In the words of A.Whitehead:

"*Philosophy is at once general and concrete, critical and appreciative of direct intuition. It is not - or, at least, should not be - a ferocious debate between irritable professors.*" - **A.Whitehead, Adventures of ideas**.

I think that it is possible from a 'third-person' point of view, to construct a coherent discourse out of and including apparently opposite views, like the ones of the realist and the illusionist about consciousness, which is closer to the reality than both of them. To illustate my point of view, let me use an analogy with picking a movie to watch at the movie theater: when making such a choice, I usually use a trick which consists in checking movies that are released in small movie theaters, usually oriented to intellectual movies, and the ones which are released in larger movie theators, oriented to popular movies; when a movie is in both lists, I am confident of its quality. I have never been disappointed by this strategy. The reason, I think, is that these movies realise a right balance between entertainment and humour, and the approach of a serious subject in a subtle way; in other words it realises a balance between two tendencies in the individual taste that is reinforced socially although it may lower the quality of the cultural content, restricting to one aspect of the reality - emotional or intellectual in the case of movies, theoretical and experimental for consciousness. Standing in between these tendencies does not mean producing a new cultural content only out of simple combinatorics of established styles, but coming back to a more original meaning, and valuing the subject as it is over what one is able to say about it or the way one is able to say something about it.

One way to compare approaches of consciousness one a comparable basis is to consider the way these approaches make judgements about consciousness, in particular why consciousness is considered as a *problem*: this is what D.Chalmers [C18] called the **meta-problem** of consciousness. To add upon D.Chalmers approach, I think this might be a more fruitful way to conceptualise consciousness than refining existing positions - especially when there is a lack of concepts to support adequately these positions. I tend to think that consciousness is a problem for its existence enters in contradiction with the way we understand the world; thus I would like to focus more on conceptualising the way we arrive at judgements about consciousness such as '*the mind works as a machine*' or '*no it does not*' or '*some being is conscious when there is something like it is to be this being*'. In particular the formulation of these different statements seem to differ in how introspection is used to speak about experience. In fact expressing these statements on a common ground can not only make a clearer sense of them but allow to push further the introspective mechanisms by which we arrive at them.

In the present text I would like to present one such way to use (creative) introspection. Although the framework that I presented in the last paper [G20] should be considered only as the beginning of a longer study, it appeared more critical to develop further the *type* of approach it consists in rather than following one approach, and put in competition the approaches that it can produce in their conceptualising power, and their capacity to build a collective discourse (which is a criterion which matters because it critically discriminated, in the XVIIth century, alchemy and the modern form of science). I would like to mention also that the particular way of thinking that I try to develop in this text imports from the area of my work in mathematics before getting interested in the notion of information integration (computability and multidimensional dynamics) a way to think about its objects towards the field of consciousness studies. As a matter of fact the intuitive idea of a relation between (un)computability and consciousness has been posited by R.Penrose, who suggests that non-computability should be an aspect of consciousness of any conscious being (see for instance his talk *Mathematics, Mind and Consciousness* [P20]). Although it is speculation, an understanding of the limit between the computable and the uncomputable in mathematics may lead to some insight in consciousness, as well as the study of faculties of the mind which are beyond any computation, such as understanding (to which I will devote another text). In fact what I think undecidability results suggest relates to the thoughts of H.Poincaré: the source of mathematical creation is not algorithmical and instead relates to the specific structure of the experiencing subject (and its relation with the unconscious).

In order to get more precise on how my (modest) work on computability can relate to the structure of the experiencing subject, let me give a short summary of this work. After A.Turing introduced his computing machines in his paper *On computable numbers, with an application to the entscheidungsproblem* in 1936, R.Berger (1966) [B66] and then R.Robinson (1971) [R71] proposed some constructions embedding computing machines inside hierarchically structured sets of plane tilings in order to disprove H.Wang's conjecture that there exists an algorithm which can decide if one can tile the plane using a finite set of decorated square tiles under the constraint of assembling rules depending on the decorations. More recently M.Hochman and T.Meyerovitch [HM10] adapted R.Robinson's construction in order to characterize the possible values of entropy that these sets of tilings generate, when considered as dynamical systems, with a computability criterion. Even more recently researchers in this field have been interested in how exactly dynamical constraints on these dynamical systems affect this kind of characterization (rendering entropy, for instance, 'more' computable). I have published with coauthors some mathematical results in this direction. What I think was interesting in these constructions is that the implementation of the computing machines had to be complexified and adopt a particular 'form' adapted to the constraints. As I was interested in cognitive sciences and the structure and organisation of the human brain, I naturally made some analogies with the living, and how dynamical constraints could explain how and why the human brain has the (information processing) structure it has (and thus how and why the human *mind* has the structure it has). Of course in this area of mathematics the dynamical constraints are adapted to mathematicaly study and thus simple and abstract enough to be tractable. When I think about the humain brain I think about more concrete and (yet) less well-defined constraints, such as for instance: integration of information (in order for instance to take into account many dispersed factors in making a decision), resistance to small perturbations, the possibility to mechanically and rapidly explore many possible decisions in order to adopt the 'right' one (this kind flexibility constraints contradicts in principle the one of information integration, so there has to be a trade-off), the possibility for a functional region of the brain to be used for various purposes (adaptability), etc. The possibility of the human brain to simulate any possible algorithm should be related to the adaptability to the evolution of the world. The result of these many constraints of 'local' and 'global' adaptability and structure 'force' the brain to be organised in its particular way.

As in my other paper [G20], one can see a tiling as the description of a (visual) experience. Furthermore one can think of an experience at a certain time (including all senses and not only visual) as described by a similar object. As well my (temporal) experience can be thought of as a series of such objects (with small enough temporal gaps). This experience is conceptualised with causal relations between 'patterns' in this experience (for instance '*if I drop an apple from my hand, it will fall on the closest object down in the vertical direction*'), collected into a structure (for instance causal relations can be chained to obtain other causal relations, and there exist elementary causal relations which can not be obtained this way). Science explains or disprove these causal relations by constructing a picture of the outside world, and its causal structure, such that this 'finer' causal structure can be projected on the causal structure of the experience (this allows the mind to act on the world using this finer causation). In its current form cognitive science tend to search in the physical an explanation for the phenomenal, while the interpretation of explanation in terms of causal structure allows other ways. Furthermore considering the set of possible experiences in which a first-personal world consits, it is reasonable to assume that it can be described by a set of rules which can be obtained algorithmically. A mechanistic interpretation or explanation of how this first personal world 'works' is a similar object which can be described by a finite set of local rules, accounting for what happens at the 'microscopic level'. The object 'first-personal' world is similar to the one of 'micro-world' that I defined in my last paper, however it is a lot more difficult to discuss directly because its extension can not be grasped (in particular because it changes) and thus to communicate and to make progress in its understanding. I think one should instead consider larger and larger 'parts' of the first-personal world which are causally isolated and map them to a progressively constituted model. In this text my purpose is to begin with the statement of A.Turing that the human mind can simulate any of his computing machines, that I take as the description of a fundamental mode of the experiencing subject, and try to complexify this model using some introspective method that I will define in this paper, using in particular theoretical constraints in terms of information transport and the account for high level phenomena. Ideas that I will propose are speculative, as only partly based on experimental evidence, but they should be seen as steps in a theoretical process which posits hypotheses before refuting them repeatedly, until one arrives at a satisfying theory. They are presented in the second part of this text. In the first part I abstract a framework for introspection which is based on classical papers that I interpret as steps in the construction of an introspective reasoning about consciousness.

II.– Elements of an introspective method

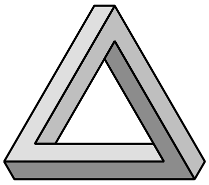
The purpose of this section is to gather elements of an introspective method in order to investigate the structure of the experiencing subject (which I will begin to do in Section **III**); it consists in analysing philosophical positions about consciousness and how they make use of introspection to conceptualise it. Before entering into this analysis, I would like to rule out arguments on introspection of its main philosophical opponent, namely illusionism, by the idea that illusionism consists ultimately on the rejection of introspection based on its misuse (including by illusionists). Comments on this positions are not the result of a thorough review on illusionism and only on a pair of recent papers of K.Frankish [F16] and D.Dennett, that I take as representing well enough illusionism for my purpose here.

II.1 – On illusionism - introspection taken as a non-method

Illusionism is, according to K.Frankish [F16], the position according to which "*phenomenal consciousness is an illusion*" and "*aims to explain why it seems to exist*". To this position is opposed the one of realism, which is that phenomenal consciousness is real. I find interesting to note that the example which is almost systematically taken to illustrate this point of view is the one of the feeling of pain: since pain exists, is specific to a conscious being (in order for pain to be felt, it has to be felt by such a conscious being) and that the feeling of pain differs from what happens in the physical world when I see such a being in pain, then phenomenal consciousness exists. This position is simpler to analyse because it does not offer several seemingly different arguments: while it seems extravagant for realists to deny the existence of pain, I think that the confusion has to be found in the meaning of the word '*existence*', and in order to enlighten the use of existence by realists I think it is interesting to wonder why the example of pain is so reccurent. The feeling of pain is in fact one example of 'object' that I can experience and which I can designate in a statical way (it does change over time, although there may be differences in how I feel the pain); this is why I can designate it to the other and that the other can recognise, and answer to the question: do you feel pain now ? In this sense of existence, it does not matter wether or not subjects can find it intersubjectively out-there.

II.1.1 – What is an illusion ?

In order to try to understand the opposite position, that is the one of illusionism, one should search for a definition of what an illusion is. For that I would like to consider one simple visual illusion which will serve as a paradigmatic example for this analysis. Such an illusion is first an experiencial situation in which one considers a bidimensional colored picture, such as the Penrose triangle (considered by N.Humphrey) which I reproduced below:



In this situation one should differenciate the picture itself, which I think as the raw experience described by an array of bits of information encoding for which color is displayed on which position on a square grid of pixels (which is stored somewhere in the computer used by the reader), from what is perceived in this picture, in other words the experience that one has of it. This experience contains for instance the 'impression' that each of sides of this 'thick triangle' are tridimensional. I think it is reasonable to say that this impression can be thought of as a conceptualisation of the raw experience described by with the array, which is 'added over' the picture itself. Thus far the situation consists thus in a picture augmented with a conceptualisation of it. Furthermore we are used to analyse (most of the time automatically) this kind of experience and think about it, especially because of its conceptualisation, as the representation of a tridimensional object 'out-there'. However with enough attention I may conclude that in this case it is not possible because this hypothesis enters in contradiction with the intuition that I hold of a tridimensional space, unless I think of two of the sides as not connected in the tridimensional space, although they coincide in the bidimensional representation (this is the idea of Gregundrum, named by N.Humphrey after its creator R.Gregory).

This kind of situation is called an **illusion** because it is deceiving. One may reformulate this and say that in contact with this situation, the mind is *distorting* what there is to perceive in it. I think however that we should be careful about what is exactly happening with this contact: **i)** my mind makes automatically a conceptualisation of the picture while experiencing it; **ii)** then I may or not make a judgement that the picture is a bidimensional representation of a tridimensional object which has some particular property (similar to the property of the representation, that the sides are two by two connected), which I make based only on the conceptualisation and a certain thought habit; **iii)** I may or not believe in this judgement. The situation is deceiving when I arrive at the third point with a belief in this judgement, for the simple reason that this judgement is false: there can not be such a tridimensional object. In other words this object does not exist, it is never perceived. On the other hand I am not really distorting the picture because I am only adding a conceptualisation over it, and what I really see is still accessible via introspection - an introspection which is not straightforward but carefully executed. Furthermore the conceptualisation and the raw experience do exist to my mind, at least as much as my actual experience because factually I am constructing them out of this actual experience. When the picture is factually a projection on a plane of the Gregundrum, I may change its orientation and observe that the tridimensional object is not the hypothetical (impossible) object; however when I come back to the initial position, this has negated my judgement but not the conceptualisation of the picture, which is still there. Moreover it does not resolve the error which underlies the factual judgement, which is that the hypothetical object is impossible. What may remove definitively this error by a reasoning of mathematical nature, which uses introspection in a non straightforward way, meaning an introspection not only about the situation in question but a set of possible situations which consists in the tridimensional completions of this one.

Looking closely at how we arrive at certain judgements - the idea of D.Chalmers - can lead to envision other possibilities, ultimately proving or disproving these judgements. This can be applied as well to judgements on consciousness itself.

I will then take an illusion to be a situation in which one is misled by an automatic conceptualisation, if not resisting to certain thought habits, into holding and believing in a false judgement about the experiencial content of this situation. I think that science, and in particular methodical science, has been constructed over the idea of a '*disillusionment*' process, but has falsely identified the observation of the material world as the only way to access '*reality*' - in particular excluding introspection from the method of inquiry.

II.1.2 – The use of illusion in illusionism - about distortion

The illusionists reasonings mainly consist in analogies between some experiencial situations which involve what realists may call 'phenomenal consciousness' and are not well understood and simple illusions, typically to conclude that there is nothing 'phenomenal' in experiencing this situation phenomenal consciousness does not exist and only 'seem' to exist. When reading the texts I found that this comparison is often superficial - in the sense that the there is no precise correspondance drawn between contents of the two situations involved - and confusing: to use the vocabulary of illusionists, these two situations 'seem' to be similar but they are not. In fact one may notice in the texts that there is usually no definition of what 'phenomenal consciousness', as well as to 'exist', is taken to be. I usually take 'phenomenal' to be in my actual experience the part of the conceptualisation of the raw experience that I perceive as received from the senses, while illusionists seem to take the term 'phenomenal' as refering to a vague feeling about the raw experience. Moreover it seems that something 'exists' for illusionists when it is present in the raw experience only, and it 'seem' to exist or to have certain properties when the object in question, or the properties, lie in the conceptualisation of the raw experience. Sometimes the term 'datum' seems to be used by illusionists to refer to what lies in the raw experience (in the very conventional sense of experimental science I believe), while realists seem to use the term 'datum' to refer to something given, which is present to the mind, and which can be taken intersubjectively as an object. This confusion appears there is the discourse of D.Dennett [D16]:

> "*Imagine Chalmers' declaration that phenomenal consciousness is a datum, transposed into the claim that a lady-sawn-in-half is a datum, or the claim that we are directly acquainted with the real presence of a lady-sawn-in-half. You may think you're directly acquainted with this, but that's a fact of personal psychology, at best an unshakable intuition, not a datum.*"

In the example of the lady-sawn-in-half, it is clear that the lady-sawn-in-half is an impression, leading to a false judgement, and not in the physical situation which is projected in my experience. So it is not a datum in the sense of D.Dennett, but the impression itself is a datum, it exists in the conceptualisation of the experience (otherwise the sentences of Dennett are empty of meaning), as well as the paradigmatic example of realists - pain - when it appears in a situation. I believe that a careful analysis of what an illusion is composed of and a more careful use of analogies may remove some misunderstandings. The same confusion happens with the comparison of N.Humphrey between sensations and impossible objects such as the hypothetical tridimensional object of which the Penrose triangle would be a representation: sensations should be compared to the conceptualisation of the picture and not the impossible object.

I think in fact there should be an explanation to the divergence between the meanings held for the term 'exist' by realists and illusionists. What I would propose for such an explanation is that illusionists attribute more value to certain parts of an experience, which seem to be the ones on which the current methodology of science can apply, where realists seem to refuse to apply such a difference in value. It appears clear to me when I read the following in K.Frankish's article:

> "*In another analogy, Rey compares our introspective lives to the experience of a child in a dark cinema who takes the cartoon creatures on screen to be real (Rey, 1992, p. 308).*"

In fact these creatures *are* real, although children may be wrong to think that there are present in the physical world (which does not necessarily make sense to them). In fact if we were about to talk about children psychology, we could not avoid considering their fictions as real objects of study (otherwise, why bother ?) and thus have them valuable. I have to recognise that the above analogy may be convincing, but only because we tend in general to deny value to the experiences of children. As I mentioned above, science has been constructed on the attribution of more value to certain experiences - in particular the observation of the material world. This valuation is often implicit and unquestioned, but lead to a real distortion of certain experiencial situations in order to make the argument more efficient, as in the following analogy of D.Dennett, quoted by K.Frankish:

> "*Dennett compares consciousness to the user illusions created by the graphical interfaces through which we control our computers (Dennett, 1991, pp. 216–20, 309–14). The icons, pointers, files, and locations displayed on a computer screen correspond in only an abstract, metaphorical way to structures within the machine, but by manipulating them in intuitive ways we can control the machine effectively, without any deeper understanding of its workings. The items that populate our introspective world have a similar status, Dennett suggests. They are metaphorical representations of real neural events, which facilitate certain kinds of mental self-manipulation but yield no deep understanding insight into the processes involved.*"

In this analogy again the two situations *seem* to be similar but are not: they have a different structure. In fact in the experience, the screen *is* the experience, and the the structure of the machine is in principle observable directly; drawing on the analogy of Dennett, one can see the interior of the machine only via a simulation of it (like a virtual machine) which is displayed on the screen - in fact in the reality it appears on a screen simulated on the actual screen. Putting it this way, it is not clear anymore that the interior of the simulated machine is more real than other objects which appear on the screen (files, pointers, icons). The only way to make sense of the fact that they are 'real' is that they are considered more valuable - and this boils down to say that they are real because Dennett says they are. One may notice also that the analogy of Dennett breaks on other points, such as the fact that while sensations exist but are not easily delineable, icons and pointers are. Moreover what the virtual machine proves is not less abstract than icons and pointers: the way we think about both of them (and unterstand them) does entirely belongs to the conceptualisation (in particular simulated mechanisms would consists in sets of causal relations, as well as pointers and icons) we make of a raw experience. So if we were to say that icons and pointers are not real, what would be ? This reminds me of Zhuangzi's celebrated *dream of the butterfly* [Giles]:

> *Once upon a time, I, Chuang Tzu, dreamt I was a butterfly; fluttering hither and thither, to all intents and purposes a butterfly. I was conscious only of following my fancies as a butterfly, and was unconscious of my individuality as a man. Suddenly I awaked, and there I lay, myself again. Now I do not know whether I was then a man dreaming I was a butterfly or whether I am now a butterfly, dreaming I am a man.*

It is difficult for Zhuangzi to decide if he is in reality a butterfly or a man, but it is difficult to conceive that he is none of them. A more reasonable position would be, in my opinion, to say that a priori both the butterfly an the man are real in a certain sense. In order to navigate individually in the experience and understand it, there is no need to distinguish between 'real' and 'not real'; at the level of the constructed collective discourse, I think this should be also the case, as long as the terms used have a clear (stable) meaning, especially if what is said can act, one way or another, towards more understanding (providing for instance the idea of a useful concept).

II.1.3 – Another way to use introspection

I think that in the term 'theory' may also refer to multiple types of discourse, and in order to think about the type of discourse, it may be useful to consider the social effect of this theory - in it may lie the real nature of the theory. In the case of illusionism, it seems that the theory is more an attempt to direct researchers to a certain method of inquiry more than about consciousness itself. It is reasonable to take a position of conservatism a priori, meaning using in a priviledged way the existing and accepted methods of inquiry, before rushing into revising these methods. However it is not clear if the application of these methods leave more conceptualisation and understanding capacity than hopes. Furthermore the manifestations of illusionism often take the form of an unjustified prompting to disengage from any other method, in particular favouring rational thinking over intuition - ultimately a progression into understanding has to involve both. The effect of this discourse may moreover be ultimately to take away even more methods that individuals possess to understand and explore intellectually without having to rely only on academic (in particular scientific) authority, an effect which may itself motivated by the factual status of the intellectual who works at understanding the physical reality.

I think that the reflection on illusion may on the other hand provide useful intellectual tools for inquiry, for they consist in situations in which the subject is in contact with a raw experience which can be completely described, whose conceptualisation can be agreed upon, and at the same time contains concepts which appear only in the experience of the subject and not out-there - which means that I can not see this concept as an actual cause of the behavior of another subject: for instance I can feel pain wihout this pain being visible to the other and possibly inducing a certain reaction of this other. Removing the negative value that is often attributed to illusion, one can see a illusions as '*toy models*' for investigating how the conscious subject constructs (unconsciously) a conceptualisation of its raw experience: the kind of explanation provided for these toy models may then be used to approach more complex situations. As the conceptualisation may be expressed and intersubjectively agreed upon, it should be possible to relate there the physical and the mental. On the other hand it is not clear if the kind of isolated experiences corresponding to illusions can lead to an insight about consciousness as such and not only about particular experiences.

I believe that the idea of illusionists that introspection can not be the source of reliable knowledge is a mistake - it is possible though that what illusionists call 'introspection' corresponds to what I call straightforward introspection. In fact I think that introspection should not be thought as a way to 'inspect' the content of experience, but to simulate the creation of concepts which make a priori sense of what appears in the mind - of course the value of particular concepts is determined by their further capacity for understanding. Along with this use of introspection, the epistemic position of the intellectual should differ from the image of a holder of knowledge to be supplied to general other as a product and resemble the one of a participant in the constitution of a collective discourse, proposing a narrative articulating concepts, from which concepts may be extracted, modified, re-articulated in other narratives, whose capacity to provide an understanding may allow them to resist natural selection. In other words, individuals do not have an authority on the inner world which would come from a natural authority on their own - on this point I agree with Dennett; only (some) concepts and narratives may eventually acquire authority. In this picture I believe that a 'scientific approach' of consciousness should not direct towards the physical world but to the constitution of narratives articulating 'elementary' - and in particular stable - concepts. In order to do so, one can try to *condensate* unstable concepts into more stable ones, by progressively concentrating the look on thiner and thiner areas until an object appears; in such a process the use of analogy is usually not enough to make sense, and can only serve as an introspective tool. Moreover the consistant look in the direction pointed at by a certain concept may make appear (in the sense that they were not present there before) more elementary concepts in the corresponding cognitive area, which constitute the designating one. This corresponds to an abstract form of increasing 'discernment' which I believe allows the use of introspection in a general sense. I thus agree with Dennett [D15] on the following to a certain extent:

> "*Qualia seem atomic to introspection, unanalyzable simples—the smell of violets, the shade of blue, the sound of an oboe—but this is clearly an effect of something like the resolution of our discernment machinery.*"

as well as the following:

> "*If our vision were as poorly spatially resolved as our olfaction, when a bird flew by, the sky would suddenly 'go all birdish,'—that peculiar, indescribable birdishness that one would experience in the visual presence of birds. And this resolution is variable: music lovers and wine enthusiasts and others can train up their ear and their palate and come to distinguish, introspectively, the combining elements of what used to seem atomic and unanalyzable.*"

Some of the judgements on phenomenal aspects of experience may reveal false, - in particular indecomposability for some of them - however I think this would be wrong to say that they do not exist, just as birds do exist even if 'in reality' they are collections of atoms - they do at least as a concept, which we may explain, if so, why it is a natural one. In fact I believe that the belief of indecomposability of pain for instance may well come from lack of introspection - which is a consequence of the simple fact taht we usually avoid the feeling of pain. When looking closer to this feeling, I personally distinguish different elementary feelings, like some 'tension' in the locus of pain, as well as some heat.

On the other hand it is difficult to conceive that colors can be decomposable, and that ultimately the decomposition of a feeling like pain could not involve 'phenomenal elements' similar to colors. Furthermore the fact that consistant introspection reveals an articulation of more elementary concepts does not mean that the initial one can be reduced to this articulation of concepts - they are a priori different experiences. It is probable on the other hand that they are '*functionally related*', meaning that there is a transformation from one to the other which is not the simple collection. In order to 'explain' the initial concept, - wether considering it a distortion or not - it is thus not sufficient to 'look closer': one also has to find how the transformation is operated. Let us consider looking at a unicolor picture for instance. The conception that we hold is that the information of this picture is processed by discrete devices - neurons - and that the real picture that we see is a discrete object; but the picture appears as a continuum: by which transformation does the brain transforms the discrete picture into a continuous one ? In fact it is not even clear that with introspection the initial concept disappears and is *replaced* by an underlying 'reality' - in particular in the case of wine enthusiasts - and it is possible that the perception is enriched without reduction. This corresponds in fact to the use of introspection that I would like to develop.

II.2 – Introspection and understanding of what consciousness *is*

While enriching the set of concepts which serve to describe the experience is necessary, and possible via introspection, it is not enough to *understand* what consciousness is. I see here two aspects of understanding which have to be intertwined with the conceptualisation itself: first the choice of a path for introspection, and then the structuration - in particular hierarchisation - of the conceptual set acquired along the way. I will explain what I mean by this in theory and then in practice, before turning to an example of their implementation in *Integrated information theory*.

II.2.1 – Foundations

Where to direct introspection? I take as a fact that the purpose of religion is and has been to conceptualise human experience in its most directly experienced aspects in order to provide means to '*navigate*' amongst possible '*forms*' of conscious experiencing - in particular to escape from a form characterized by being in a place called *hell*; or if you are reluctant to use this term, you may think about the inner existence of R.Raskolnikov, in *Crime and Punishment* [**F.Dostoïevski**], after his murder of the old pawnbroker. Several philosophers have discussed the '*why*' and '*how*' of this kind of conceptualisation. My point here is not discuss highly controversial subjects but to ask the following question: how can a scientific approach differ from religion on the matter of consciousness and what can it bring to an understanding of this phenomenon ?

I believe that it can differ in finding 'mechanical roots' for conscious experiencing, rooting or disproving statements about it (or practices which act on it), which may make them more accessible to understand and rely on. I think that in doing so one does not have to rely (only) on the study of the physical world and the structure of the human brain, but also on a use of introspection which should progressively reveal 'elementary' concepts - which do not have to be necessarily of physical nature - of experiencing and their articulations. Understanding the experiencing subject here should mean conceptualising its structure in terms of these elementary concepts in such a way that the effect of a targeted perturbation (by the subject himself or herself) can be conceived and predicted - a prediction which in principle does not have to be mechanical. In praticular the concepts themselves should be *statical* designations, meaning that what is pointed out by the concept in the experience does not change.

I think that such an understanding should be done before any attempt of defining an external measure of 'consciousness'. Furthermore I do not believe that consciousness could be measured as a quantity, but as a specific *type* of activity pattern (that is, not necessarily computation in the usual sense), which is shaped for instance by teleological constraints (the simple subsistance), and which indicates potential *effective* consciousness, meaning the possibility for a person to interact with others, one way or another, involving the acquisition of information, the possibility to emmit a judgement about this information, about its interest, etc. The so-called '*hard problem of consciousness*' may philosophically be interesting to investigate aftermath, but it is detached in principle from practical applications.

Introspecting on the structure of experiencing, I think that '*where*' I should look at is where the conceptualisation out of statistical designations may be possible. I think that looking at the dynamical relations between the conscious mind and the unconscious one, or in other words how objects appear and disappear along with the activity of the mind should reveal something about experiencing more than what constantly appears to the mind - for this is in principle accidental. Moreover the understanding of the structure of the experiencing subject should involve a 'local' study, meaning the *type* of experiencing dynamics (for instance access to an information, or active integration of multiple informations, etc) that the subject can have in a particular cognitive area, as well as a 'global' one, meaning how these types of dynamics are articulated by the subject - probably hierarchically.

How to structure the conceptualisation? Conceptualising is at the beginning enriching the set of concepts available in the particular cognitive area that is looked at. However some of them may not be useful and in fact potentially obstructing understanding - since it is difficult to deal with and manipulate too much informational content simultaneaously. In order to arrive at such undersanding, one should then select some more important concepts over others and structure the remaining ones. Each intellectual practice has its own way to make this selection and structuration, in a way that is adapted to the objects it is dealing with, and which is manifested ultimately in the language structures which compose the corresponding discourse. As a matter of fact, approaching consciousness scientifically, it should not be expected that the discourse about it should adopt the structures of languages which compose the discourse of science in general (as I mentioned in Section **I**), in particular because the 'object' of study is not extracted from an experience or a set of experiences, but is relative to experiencing itself. For this reason we should retract to the experiencial foundation of the methods of science before re-deriving and synchronise intersubjectively around adapted language structures. In fact the fundamental idea that I take from M.Merleau-Ponty of how science regulates concepts is by negation of negation. By the first negation I mean the act of the subject to try to suppress transcendentally a certain idea. Let us take for instance the idea of truth of a particular formulated theorem: I may negate this idea by holding a certain notion of what is a mathematical and attempting to detect in the proof of the theorem some error which corresponds to this notion. If I do then I have to abandon the idea (at least momentarily) that the theorem is actually true. Otherwise if after some time I do not find an error, then my negation is negated (second occurence of negation). In fact what science - here I include mathematics - holds as true is what resists negation. This idea is realised experiencially in different ways, according to the field, even outside of science:

why for instance do we want solipsism not to hold ? There is no logical or physical reason for that other than than the emotional negation which consists in the extreme feeling of solitude. In principle, without this negation, solipsism seems to be a perfectly valid position from an individual point of view. If one accepts that solipsism does not hold, then the experience of the others exist 'somewhere', which means that it is necessary to re-think our understanding of space and time for this somewhere to make sense - I believe that it is on this premise that C.List's *Many worlds theory of consciousness* [L21] is based. I think that drawing by introspection a 'map' of feelings of distance and proximity to the others may lead to an idea of how to formulate more precisely this. I usually also consider that concepts can be regulated simply by forgetting: if a concept is not recurrent in the mind, triggered again by a question or an observation, it is unconsciously considered as meaningless and forgotten - maybe this could actually be a definition of meaninglessness. Only central concepts of the situation considered are left with time, and the 'degree' of recurrence should provide a mean to structure and hierarchize these ones.

II.2.2 – In practice

In practice some cognitive areas are more difficult to introspect than others. The notion of '*information integration*' is an example. I may have many ideas of how to picture it with articulated elementary concepts, or in other words how it is 'realised' in my mind, or how it may be realised in the human brain, but there is no clear way to select one over the others, and no clear way to grasp a '*space*' of all the possibilities to picture it. In fact in local cases I can actively integrate two informations when I am planning for instance: '*when the clock will display 9:00, I will go out*', and I can picture that these two informations are connected for some time by a 'link' that I create between the two. However it is not possible for me to picture straightforwardly how different aspects (color, shape, for instance) of the same object are integrated, or how the 'I' and the phenomenal experience are (for I can say that this experience is 'mine'). Of course integration may be realised in different ways but if I could grasp all the situations of information integration at once - just like we can grasp all the possible triangles in a single concept - I could find what is common in all information integration phenomena and possibly derive and study the particular from the general.

There is no other way to find out if a certain area of cognition is '*introspectable*' or not than to try, which means focusing the attention towards this area and see what appears there, and with practice focusing on areas where original concepts appear, which moreover may be useful once fully formed in other situations, in particular ones where introspection is difficult. I found useful, when a cognitive area is introspectable but no clear definition of a concept appears, to put artificially in '*competition*' possibilities of conceptualisations by creating a framework which allows the systematisation of this competition and the record of the significance of each idea - this allows in particular to put in competition my ideas with ones of others - with the aim that one or few of these ideas appear progressively to be most significant. This was one of the purposes of the last paper [G20].

Several accounts of 'consciousness' using mathematical language have been proposed, including *Integrated information theory* and *Predictive coding*, as well as - to a certain extent - *Global neuronal workspace theory*. I believe that this can be relatively surprising, and it is natural to question the use of mathematics for conceptualising consciousness as such. I believe that the nature of the project of understanding consciousness in mechanistical terms, that is finding elementary concepts and their relations to describe it, calls for the use of mathematical language: indeed, when sustained introspection is necessary to make these elementary concepts appear to my mind, I expect that the situation is the same for others. Since the corresponding designations are 'thiner' than the ones of the common language, I may not be able to use words to describe the content of my mind after the sustained introspection. Instead I could designate abstract objects and the relations between them using mathematical symbols - since they are a priori semantically neutral, they immediately refer to '*an abstract object*' - with the idea that this structure, when projected on the cognitive area in question, will reveal to the other relatively unambiguously the concepts I have in mind. In order to explain how this might work, I like to think about Sperling experiment: when my mind is clear, an unconscious content does not appear because it is in competition with many other contents, but when I hold a certain object in mind - in particular some abstract structure - the unconscious contents which are somehow related to this object appear (they are selected).

However the use of mathematics in order to describe the mind in general is very delicate. The main reason is that we tend to assimilate mathematical objects to the symbols that represent them - an idea that I expressed in [G20], while they refer to a mental reality (although stable), and often times when a mathematical symbolism is used for another purpose than its initial one, the two mental realities are assimilated, and meaning is lost along the development of the discourse. One of the rare occasions which led to a meaningful formalism was the formalisation of computation by A.Turing - a meaning partly due to the Church-Turing thesis, in particular that the mind can '*simulate*' any possible computing machine. The method that I propose in Section **III** consists in reconsidering the cognitive situation in which this statement appears, and progressively account for aspects of the conceptualisation of experience in situations which are closely related to this one, in order to propose ultimately a 'model' for the structure of the experiencing subject. The remainder of Section **II** will be devoted to integrate in this study some principles of introspection which underly intuitions of philosophers about consciousness, in particular in relation with computing machines; this is also the occasion to rule out too radical positions concerning the idea that '*the mind works as a computing machine*'.

II.2.3 – The example of integrated information theory

Before moving further, I would like to comment briefly on *Integrated information theory* and its use of introspection and mathematical language, providing a complement of my critics written in [G20]. The text [IIT] offers an overview of this theory; I also recommend [KT20] for an exposition of the formalism aimed at researchers with formal training. The theory begins with introspection of fundamental properties of experience in general - what proponents call *axioms*; here introspection is straightforward, although the exact meaning of these fundamental properties is not clear (I proposed an interpretation that I find clearer, although not formal, in [G20]), preventing any '*verification*' that my own experience has these properties. From my point of view, the formalism of Integrated information theory is in fact obtained by projection in a fixed formal context, in other words it is expected that these properties should be expressed in a certain type of formalism - this is one of the expections I mentioned earlier that should be removed. By this shift the formalism leaves completely the experience - it is *unsituated* - and it should not be expected to explain anything about it. As a matter of fact one can find easily language constructions which have no meaning - in other words they do not point at anything in the experience, it is an empty designation - such as "a square which is also a circle". Since the formalism does not point systematically at something which is tightly related to its origin, the probability is low that it will be related to it in the end. This kind of relation has been attempted, under the form of an 'explanation' of aspects of experience (for instance in [HT19] for phenomenal space), but it is significant to notice that this kind of explanation is based on directly accessible aspects of experience, for which there are many other possible explanations, including simpler and intuitive ones. For instance I would tend to characterize space with a cognitive area '*equipped*' with an exploration process. The idea that the cerebellum does not support conscious experience is not a fact and is only a judgement potentially false (have we tested all the possible situations ? it may be that it does support conscious only in some of them). Even if it is the case, one could formulate this saying that 'I' do not have access to it, or with a teological reason that if I could have, I may not be able to control this access and disturb the functioning of my body. On the contrary, I think that an 'explanation' of consciousness should 'unveil' aspects of it that are not yet visible, and that unveiling may shed some light on visible aspects. For any formalism, or simply an articulation of elementary concepts to be meaningful, it should draw an 'introspective path' which brings the reader to a situation in which there is no already commonly accessible conceptualisation, and proposing one that captures 'what it is like' to be in this situation, and by comparison to the way the reader naturally conceptualises it, this one can confirm or refute that the proposed conceptualisation is partial or faithful.

In the following I would like to convince the reader that this is possible by a form of creative introspection which takes as basis the idea that the human mind can simulate any computing machine instead of the '*axioms*' of integrated information theory.

II.3 – T.Nagel

II.3.1 – What is it like to be a bat?

In his celebrated article *What is it like to be a bat ?* [N94], T.Nagel proposed an intuitive definition of 'consciousness', under the form of a principled criterion for an organism to be conscious or not:

> *"But no matter how the form may vary, the fact that an organism has conscious experience at all means, basically, that there is something it is like to be that organism."*

Although not unanimously accepted, this definition is often used, in particular to formulate the hard problem of consciousness: *how to explain what it is like to have a certain experience* ? which usually comes with the idea that this problem is not reachable by the scientific method (in its current form). If not rigorously defining consciousness as such, I believe that the formula of T.Nagel characterises accurately the organisms to which we attribute a form of consciousness after a careful introspection. Here I use the term 'careful introspection' for the reason that one can see - as M.Grazziano did - that we may tend attribute unconsciously conscious experience to a systems which are not likely to be so - a ventriloquist's puppet for instance. I see this form of attribution of conscious experience as a conceptualisation of the raw experience - which is similar to the attribution of tri-dimensionality to the sides of Penrose triangle. With an exhaustive examination of this experience one can picture that the ventriloquist is the cause of the puppet's movements and words and one then recognizes the illusory character to the attribution of conscious experience to the puppet. As a matter of fact, T.Nagel's formula rules out this case, for there is nothing it is like to be a ventriloquist's puppet.

The observation of M.Grazziano is the basis of his *Attention schema theory* [G15] of consciousness, which is summarized as follows:

> *"In a nutshell, the theory proposes that subjective awareness is the brain's internal model of the process of attention."*

What is common to the objects which we conceptualise as 'conscious' - or more accurately 'aware' - is that they exhibit some dynamics of sensibility (in a sense that the object reacts specifically) towards the presence of other objects that I perceive, which resembles the process of direction of attention that I execute towards these objects. The idea of this theory is thus that the illusory attribution (and not necessarily the attribution of consciousness after careful introspection) of conscious awareness coincides with the perception this type of dynamics. I think that this attribution works as a hypothesis produced automatically by the human brain in order to react rapidly to the potential presence of other animals (which may be dangerous); a hypothesis which *explains* the perceived dynamics of sensibility, and is projected onto the organism and exists in it in a similar way as the center of gravity exists in a massive object. The fact that ultimately conscious experience is not attributed to the puppet may come from other factors which are examined after consideration of the hypothetically conscious object, such as the property of *autonomy* - the causes of the object's behavior systematically have a part which lies within the object. It can also be the fact that it is not possible to augment the objects in my experience to which the organism is specifically sensible, by designating an object it is not specifically sensible to already.

The kind of organisms whose conceptualisation as 'conscious' resists to this analysis are ones for which I can conceive that there should be something it is like to be the organism. However the conception that an organism has a conscious experience may come after observation of various experiences in which this organism appears: for instance I can imagine a bacteria-like organism which reacts *specifically* to objects that I identify in experiences I have in which it appears, and I will naturally 'explain' to myself this behavior as the one of a conscious organism [this, by the way, may be an argument in favour of a mild panpsychism].

The difference between the illusory attribution of conscious experience and the attribution according to T.Nagel comes after this careful investigation of an experience or a set of experiences in which the organism appears. The organisms for which I can conceive that there is something it is like to be it are the ones for which I can, out of the observation of the organism with its 'environment', find a 'transcendental formula' (a sequence of transcendental operations) to apply on my own way to interact with my personal world in order to obtain the kind of experiences that I hypothesize it has (which is not necessarily the kind it actually has).

After this definition T.Nagel proposes in his paper to consider the case of the bat; it is possible to doubt that bats do not have conscious experience because this experience is not as such perceptible to any human. However if one does, there is no reason not to doubt in the same way that other human beings have conscious experience. Ruling out solipsism, one has to accept that bats have conscious experience. T.Nagel chose the example of the bat for the following reason:

> *"I have said that the essence of the belief that bats have experience is that there is something that it is like to be a bat. [...] But bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine."*

The experience of the bat as such thus poses a problem not only to science but human thoughts in general to access things-in-themselves, as they 'really' are.

II.3.2 – Introspective exploration

It is clear that one can not access to the the kind experiences that the bat has as such, and only as human beings can imagine it, through a series of transcendental operations on the human way of experiencing [as T.Nagel formulate these, *addition*, *substraction*, or combinations of them]. On the other hand this case can serve as an example of how one may attempt to conceive how experiencing as a bat may be; for that one can not restrict oneself to straightforward introspection. For instance I may concieve the formula that the bat does not experience visually, however by picturing the spatiality of the environment out of sounds that it produces and 'hear' back, suggesting a series of transformation on my own way of experiencing. For instance I can try to blind myself and use only my voice to see how I may represent myself the spatiality of my environment - of course the conceptualisation of this experiencing should take time. Such a conceptualisation may be extrapolated to the bat and as a hypothesis be confronted to the observable behavior of the bat.

As I describe it, this kind of exploration may only provide subjective and not objective results. However any objective knowledge can come out only of first exploring and experiencing; this exploration should be considered as a first step, and objectivity should be constructed intersubjectively with time. In fact this exploration should come with a revision of the notion of objectivity itself: as T.Nagel puts it, "*It may be more accurate to think of objectivity as a direction in which the understanding can travel.*". There is no other way, for the notion of objectivity that we collectively currently hold is that which in the experience is not relative to any viewpoint, and because the object of study is here the viewpoint itself, "*any shift to greater objectivity -that is, less attachment to a specific viewpoint-does not take us nearer to the real nature of the phenomenon: it takes us farther away from it.*". For T.Nagel this pursuit of a "*more objective understanding of the mental in its own right*" may force us to set aside temporarily the relation between the mind and the brain. I think that this does not have to be the case, but we should see this relation not as something to construct yet but a tool to direct one's introspection (I shall make this more precise in the following).

On the other hand I believe that conceiving how bats in particular are experiencing is not enough to understand the 'consciousness' that we may have in common; considering consciousness through the spectrum of the structure of the experiencing subject, the method of exploration should be to consider all the possible forms that consciousness may take - in other words all the possible structures. Understanding consciousness would be to understand this space of possibilities, for consciousness as such is what is left after removing the form of it. By way of comparison, religion takes you to explore certain forms of consciousness, in order to reach one particular form accessible to human beings; its purpose is not to explore the space of possibilities in a systematic manner, and it may be dogmatic in the way it leads to this particular form. How to explore this space ? Let me make an analogy here: if I ask you to close your eyes and give you an object asking you to tell me what shape it has - let us assume it is a torus for the example - I am guessing you will take the object in your hands and touch it on one particular place and then move the hand around this place, progressively mapping the feel of touch to a visual map, until you find this representation 'closed', meaning that for a time subjectively long enough, you always come back, according to the constructed representation, to a place you already have been to. With this representation in mind you can tell me confidently that the shape is a torus. By contrast if you had touched the object on one point, and then another far from this one, etc, it might be difficult to arrive at a faithful representation of the shape. I think it may be difficult even to direct your finger to a point of the object, so why not beginning with the one you hold it through ? - what corresponds to this point is the object of the following section.

The way a bat experiences represents only one point of this shape. T.Nagel also suggests near the end of his paper some speculations on how to arrive at conceptualizing other forms of experiencing beginning with humans (for it should be easier I guess). For instance: *how could a deaf person explain a blind one how it feels to hear sounds* ? An answer to such a question can serve as a criterion for a right conceptualization. However I do not see clearly that the difficulty would be reduced. One may generalize this approach and try to conceptualise systematically how human beings which live in differents 'worlds' can explain what it is like to live in their 'world'; however I believe that in this study one can not reasonably expect to come down to elementary concepts: this will be the object of an other article.

Exploring the shape, and conceptualising 'patch by patch', a possible way to conceptualise consciousness as such should appear as the set of invariants of the conceptualisation of patches.

Digression: *Thinking about the type of speculations that T.Nagel proposed, I thought about the exclusion 'axiom' of integrated information theory. As a matter of fact the theoretical significance of this 'axiom' is related to the dimension of the set of possible 'geometrical characteristics' for the support of an experience, as it tells that the experience could be more extended or less extended than it actually is. This significance shrinks when one considers considering slightly more abstract 'geometries' for the visual experience for instance. Let us consider the following question: what does it feel like to have a visual field having the shape of a sphere ? As a matter of fact it appears quite difficult to imagine (why is that so ?). Physically I could imagine that I have eyes all around my head, that the perceptive fields of these eyes overlap leaving no whole and all together form a sphere, and that the direction of my attention can move from one eye to the other; however I can not imagine that these partial vision fields are given all at once. I can only discover the topology of my whole vision field through experience itself and causal considerations about its content: when I go around the sphere and come back to the same point, the way I know it is the same point as the one I departed from is that what I see at these two points is always the same. On the other hand I can easily construct a 'spot-like structure' like the one computed in [HT19] which is supported by a sphere; this shows (if it was needed) that this 'structure' is not essential to the phenomenology of space.*

II.4 – A.Turing

I shall begin the exploration on a form of experiencing which corresponds to the definition of A.Turing's computing machines. The reason for this is the fact that the human mind can simulate any possible computing machine is apprehensible formally and intersubjectively, as well as that it '*approximate*' in a certain sense how the mind works.

Here again I would like to rule out some preconceptions and fast critics on how computing machines may be used in order to progress in the understanding of the mind. I would like, to begin with, to detach the approach I am defining in this paper from the idea that one can understand the mind in terms of machines as we usually conceive them - in particular computers. As a matter of fact, even such a machine is a finite collection of causal relations (corresponding to logical gates) to which is reduced a physical system constructed to support in a stable way these causal relations. I think that it is possible to study the mind introspectively and reveal how the mind works *mechanistically*, meaning that it can be described in terms of causal relations, but there is not reason to think that these mechanics can be realised by a *finite* machine - a finite collection of causal relations. Furthermore I do not think that it is possible with current understanding of the mind and the brain, to assume that the way we experience the world can be exhaustively described based on the mechanics of the brain - in other words we can not assume that causation in the mind can be reduced to causation related to a certain substance in this mind (the material). I fact I will take computing machines as a '*fundamental mode*' of how the mind works without any hypothesis on how far this fundamental mode is from the general structure of the experiencing subject.

II.4.1 – Computing machines and the human mind

As a matter of fact it matters at this point to come back to the point of view of A.Turing in his paper *Computing machinery and intelligence* and see how it can be tweaked in order to understand better the mind in its own right. The general idea of A.Turing back then was about the possibility to answer the question: *can machines think ?* - a question which would lead to a more precise definition of what thinking is. He proposed that if we can ever build a machine which can interact intellectually with a human as if it was a human from the point of view of the human, then this machine actually thinks.

One can find the first definition of computing machines in Turing's article *On Computable Numbers, with an Application to the Entscheidungsproblem.* One may notice that Turing takes these machines as a faithful representation of the mind of "*a man in the process of computing a real number*". This is convincing for the mind in the process of computing, in particular for the human memory in this process is limited (here memory refers to what the man in question can hold consciously in mind during the process), which corresponds to the finiteness of the set of possible states for the machine. Furthermore one can see the tape of the machine as the simplification of the set of paper sheets that the man in the process of computing uses to compute, which consists in reading and writing on these sheets symbols according to certain finite set of rules, which may be written on the first sheet of paper - it is clear that this simplification does not affect the way computation is done. In this article his notion of computing machines is simplified to fit the mathematical analysis, while in the other article *Computing machinery and intelligence*, he provides a way to think about how the machine may be decomposed to fit the description of how the mind actually works in the process of computing: the machine is there decomposed into a '*store*', where information is stored and consists in the tape and a part of memory in the mind, an '*executive unit*', which executes serially individual operations, and the '*control*', which ensures that the operations are executed correctly by the executive unit.

With a minute of introspection I can see that this description is correct. In a sense it was possible for Turing to transcript '*how the mind works*' in mathematical terms (causal or functional if you will) in this precise situation of the computing process. However it is not clear that the functioning of the whole mind can ever be accurately described as the functioning of a computing machine, which is the view that Turing defended in *Computing machinery and intelligence* [Turing]. This idea has resisted and still resists (in the philosophy of D.Dennett for instance), for various reasons amongst which I think the effectivity of the concept of computing machines in mathematics played an important role. I believe that the most robust philosophical argument (although negative) for his view is the counter-argument to the **(4)** *Argument from consciousness* (page 14). The argument is basically the following: machines can not reproduce certain human behaviors which require consciousness, such as writing a sonnet because of the thoughts and emotions felt, and more straightforwardly, a machine can not feel anything. It is interesting to note that the response of Turing displays ideas similar to the ones of the later article of T.Nagel. His counter argument uses the rejection of solipsism: if a machine is built which reproduces human behaviour exactly, there is no reason to doubt that the machine is conscious without doubting that other humans are. This argument is effective but relies on what I would like to call the *realization hypothesis*, that such a machine can ever be built. In particular it should also reproduce not only human behaviour in a restricted context but also human *existence*, otherwise the argument breaks as there still lies the difference between the illusory consciousness attributed to the ventriloquist's puppet and the one attributed to other humans. It is clear that Turing was aware of this flaw in the argument, however his purpose was to question what should be done in the meantime:

> "*The only really satisfactory support [...] will be provided by waiting for the end of the century and then doing the experiment described. But what can we say in the meantime ? What steps should be taken now if the experiment is to be successful ?*".

Near the end of his paper he provides more positive arguments for his view, developping the idea of *learning machines*, which can be taught by humans in order to pass the test. This led to the emergence of the field of *machine learning* and its multiple branches; despite its achievements, however, machine learning does not enable machines to pass Turing's test yet. My belief is that machine learning can be considered as a good tool to program performative algorithms efficiently, but it is far from able to reproduce human behavior exactly. As a matter of fact even if a machine does realize a certain work more efficiently than humans this does not make it closer to human being as it actually is, but farther. Turing responded this argument with the idea of introducing artificial errors in the rules followed by the machine, but the way the machine does errors, notice and correct them does also matter. Also I believe that an important barrier to overcome is the one of *heteronomy*: machines can not find by themselves the means of their subsistance and use them without a human.

I think that the other arguments of Turing can be deconstructed as well, however the point I would like to make here is that maybe the most efficient way even to reproduce the human mind from the outside is to understand how it is from the inside (introspection). For this the process itself of modeling a part of how the mind works in the particular situation of computing on computing machines can serve in order to extend this modeling onto a priori different objects, which can take into account for instance the way the human mind does mechanical operations which actual machines can not do: for instance the constant re-definition of its rules of conduct, mentioned by Turing in the point **(8)** *The argument from informality of behaviour* (page 21). In fact the counter argument of Turing in this part against the conclusion that humans can not be machines is that rules of contuct should be differenciated from 'laws of behaviors': there might actually be such mechanical laws which root the definition of humans' redefinition of rules of conduct. It is not clear however that these laws can be reproduced by machines as we currently conceive them.

II.4.2 – Non-reductive mathematics of the experiencing subject

The description that Turing has made of the human mind in the process of computing as a finite mechanical process (the computing machine) and speculatively extended to the whole mind has played, I believe, an important role in the collective adoption of *physicalism* - which I take to be defined as the idea that the mind can be described completely in physical terms, that is in terms only of causal relations between abstract objects (which are experiencial objects substracted from a priori inessential phenomenal aspects) which may eventually be represented to the mind by itself via phenomenal experience. Although the article of T.Nagel is a strong attack against physicalism, I believe that the following sentence is a rather convincing *debunking argument* against it:

"*A physical scientist does not introduce awareness (sensation or perception) into his theories, and having thus removed the mind from nature, he cannot expect to find it there.*" - **E.Schrödinger** (1958).

For the physical scientist to neglect a priori inessential aspects of experience - inessential relative to a practical use of an understanding of reality - has been natural, for the conceptualisation of a priori essential aspects of this experience has been made possible precisely by the focalisation resulting from neglecting. However it is this precisely which led to the separation of the mental and the physical (dualism, and with it the difficulty of thinking causation between mental and physical [K20]), and furthermore physicalism. I think that in order to understand the structure of the experiencing subject, it is not possible to set a priori that any of the substances in the mind is ontologically prior to the others - in other words we should include back inessential aspects in the theory and rethink causal relations between essential and inessential aspects of reality. We will see both types of aspects as coexisting in the reality - if I am interpreting well, this is similar to what C.List proposes in his many-worlds theory of consciousness [L21], that first-personal worlds coexists in a common reality. I do not think that positing this coexistence consists in avoiding the question of how the mental appears out of the physical: after all, the problem itself is generated by physicalism. Moreover we do not really have a response for how the physical world appears from the void. Why should it be different for the mental ?

In fact it is relatively simple to see how causation can happen between mental events of different natures: when I instanciate in my mind an algorithm, let's say "*walking in the streets of Kraków*", I hold in my mind (most of the time unconsciously once I have learnt these rules of conduct) the following rules: i) when encountering a road on my way, look at the traffic light on the other side. ii) if the light is green, cross the road iii) else, wait until the light becomes green iv) when so, cross the road. In this context it is fair to say that the mental event "red traffic light" is 'a' cause of me stopping. On the same model one can imagine that similar causal relations as 'laws of behavior' instead of rules of conduct. This simple example resembles to the computing machine (there is a store, an executive unit, a control), with a clear difference from the formalisation of the computing machine: while the later consists in a finite collection of causal relations between events which are abstracted from actual mental events to keep only the idea that they are all different from each other, the former consists in a finite collection of causal relations between *actual* mental events. I shall call the former a **mental computing machine**. It has been natural for A.Turing to extract from it the causal relations to form a mathematical model of computing, for they are the only essential aspects for computational *power*; the simplified version of the computing machine can be also put in relation with other mathematical objects and integrated in the mathematical discourse (in particular with the development of computability theory). Furthermore for externalizing a certain mental algorithm, it is natural to extract from it only causal relations, for this way it is possible to replace mental events with events that are possible to control, ensuring the stability of the algorithm's execution. The fact that these causal relations are realised in my experience by other subjects is also a strong factor for the attribution of meaning to this formalisation (in a very similar way to M.Graziano's attention schema). However if we are to understand the mechanics of the mental in its own right, we should reverse the operation of extraction and base this understanding on mental computing machines.

On the basis of this ontological equalisation, I believe it is possible to approach mathematically the study of the experiencing subject; not in the sense of an application of mathematical formalism, but in developping a mathematical reasoning from the psychological foundations of the practice of mathematics to experiencing itself - I shall make clearer the terms 'psychological foundations' later with the reading of J.Hadamard. As Q.Meillassoux expresses it in his book, *After finitude: an essay on the necessity of contingency* [Meillassoux], the change that Galileo has operated in the use of mathematics compared to his predecessors was not to find mathematical patterns in the experience, but to think that the description of certain experiences (the movement of massive bodies) can be '*exhausted*' mathematically - meaning that every aspect of such experience can enter as a part of a mathematical model of it. I understand this '*exhaustion*' in the same way as the term exhaustion I used in my previous article [G20], meaning that a delimited experience can be thought as the assembling of distinguished objects that it contains, in some spatial relations and causal relations (including rules of evolution). I think that in order to study the structure of the experiencing subject, one has to make a similar extension of the operability of mathematics. On the other hand I do not believe that it is clear yet what kind of formalisation this should lead to, which is the reason why I like to try multiple ways to do so at this time. Concerning what in experiencing the operability should be extended on, I think that should be first the set of transcendental operations through which the subject relate to experience (navigating 'in' it, designating and isolating elements of this experience, transforming imaginatively contingent aspects of these elements and connecting them in the instanciation of a mental algorithm, etc), as well as the structure of the "*mental space*" [K20] - for which the partial knowledge of the structure of the brain can serve as a tool for introspection, providing hypotheses on this structure that can be verified or falsified introspectively. In this direction, I think that the particular architecture of processing of information in the human brain should be the result of contraints (I believe of dynamical nature) applied on the living compared to machines (spatial constraints as well as autonomy). Similar constraints as well as the information mechanisms that they result in (I think in particular that emotional responses serve the situated focalisation on the body's imperatives and are included in overarching attention strategy) applied to the mind functioning as a computing machine may also serve as a tool for introspection.

Focusing on how mental computing machines are instanciated and operate should enable the introspective exhaustion of particular situations of experiencing, in particular how transcendental operations are realised in these situations and their spatiality 'in' the mind - I see this focalisation as similar to the focalisation operated by physics on a priori essential aspects of reality. We could see the realizability of computing machines by the human mind as analogical to an *axiom*, for it has a similar foundation role. However I would like to deviate from this analogy, for the phenomenological use of it is different: I think of it more as an introspective 'gateway'.

I should insist here on the idea that for understanding how mental computing machines operate in the mind, one has to consider how they are in the mind, "*coming back to things themselves*" (phenomenology), where here the things in question are the computing machines. By way of comparison, absurdities in logics (Russel's paradox for instance) appeared when the mental act of collection was removed from the definition of a set, and disappeared when considering it back (no act of collection can construct Russell's 'set', it is only a language construction similar to the "square which is a circle"). We should come back similarly to how precisely (by which transcendental operations) the computing machines are assembled and executed in the mind - explaining in a sense how A.Turing introspectively arrived at his model.

Digression: *I think that it is interesting to wonder why there is a difference in the attribution of conscious experience to a computing machine observed while functioning (no attribution) and other humans or animals (attribution), while they both exhibit an attention schema. I think that this means there should be an additional factor than the attention schema which is essential to the attribution of conscious experience. I think that one factor is what I mentioned earlier: the specific sensibility to objects to which I am specifically sensible. I shall add that these objects should be 'close' to me in the sense that the sensibility is not artificially constructed (such as the symbols the computing machine 'pays attention' to).*

II.5 – J.Searle

I think that the above digression should shed some light on J.Searle's *Chinese room argument* [Searle], which has an important role in the introspective method that I am searching to develop, although not as an argument but as an introspective technique.

This argument is meant, roughly, to convince that '*instantiating a computer program is never by itself a sufficient condition of intentionality*'; one consequence posited by J.Searle of this argument is that '*any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain*'. This consequence can be extended a forciori to replace intentionality with 'consciousness' - meaning essential aspects of consciousness as such.

This argument goes roughly like this: imagine there is a program which can, provided a text written in English, translate it in Chinese (here replace this language by any one you don't understand at all, let's say Swahili for instance), in such a way that the translation is convincing for persons who speak the two languages fluently. Imagine then that you are in a closed room and assigned to execute the program step by step on a text provided by persons out of the room through a hole and then output the obtained text in Chinese through another hole. It seems clear there that although you don't understand Chinese, the room as perceived from the outside passes the test for the specific assignement of translating a text in English to a text in Chinese.

J.Searle's conclusion is that although such a machine can act as if it understood Chinese language, it does not imply necessarily that it does understand it. For J.Searle, the attribution of some faculties - such as understanding, also posited by R.Penrose [P20] as a difference between minds and machines as we conceive them for the same reason as J.Searle that it lies beyond the execution of a finite set of rules - comes only from the tendency of the mind to project onto machines which exhibit similar observable behaviors which are only loosely related to these faculties. In other words, in this situation the attribution of understanding (as well as other faculties of the mind related to intentionality) is illusory.

I think that computing machines (I believe used implicitely by J.Searle) are particularly suitable for this argument because it is possible to identify a point of view 'in' the machine (the machine head which is moving on the tape), from which I can stand to imagine *what it is like* to be such a machine - or more precisely 'what it should, or could, be like' (once the realisability hypothesis is supposed to be satisfied) - otherwise it would be difficult to picture this out of my own experience. After taking the position of this point of view, I am able to compare the conception I can form of *what it is like* to be the computing machine, which in fact is a structure of experiencing (and not a particular experience or series of experiences), and the structure of my experiencing. It is this introspective comparison which renders more present to my mind some elements of the difference between these two structures of experiencing. I believe that J.Searle chose understanding for the imediacy of the argument, but this exposes it to the critics because 'understanding' (he uses also the notion of 'belief') - as he recognizes it himself - does not refer to a precise notion. I believe that the simple presence to the mind of an object to which the words refer, as it is required for understanding or beliefs, would be sufficient: I do not understand the Chinese symbols because I do not have in mind any object it refers to. In fact when someone translates a language from another, this person uses a mental machinery which to each word in the text to be translated associates a mental content and make these contents present to the mind altogether for the words of a sentence respecting its structure, in order to arrive at the representation of a situation. Then this person can find words in the other language in order to describe exhaustively this situation and assemble them in a sentence respecting the structure of the mental situation. It is this mental machinery which is suppressed in the conception of translation programs. Furthermore it is conceivable to design a machine which 'understands' Chinese in refined sense - for instance it could be able to answer questions on the individual symbols and words that it manipulates, just as the questions a teacher asks the student to verify they understood - but there will still be missing the presence of a reference. Besides, the computing machine model can not render the variations in the presence of an object to the mind (for instance the discovery of this object, meaning the perception of an object which I can not identify with another object I can conceive or extract from a past experience) or the possible causal interactions between the conscious and the unconscious.

I would like to see this introspective comparison as a tool in the process of progressive '*approximation*' of the human structure of experiencing by finite mechanistical accounts. There is no reason a priori to think that this approximation process can exhaust its object in finite time - this is the obstacle of potential *infiniteness* posited by R.Penrose [P20], and mentioned earlier by A.Turing [Turing] as a potentiality. The reason is that a better approximation may lead to more accurate introspection, for instance the presence of an element in the difference between the two structures of experiencing (to be still conceptualised) that was not perceptible earlier in the process.

J.Searle's argument received several critics, some of which can be found in [Searle]. I would like to spend some time here on analysing some critics in order to defend its use in an introspective method, although it does not meet some criteria of definiteness. I recommend reading the one of R.P.Abelson (page 8) for it is quite amusing proof that the name Yale is not a synonym of brightness - also, I hold as a hypothesis that some "scientists" may identify themselves so much with machines that they need them, in order to preserve their ego, to have the faculty of understanding. More seriously, the one of D.Dennett (page 12), which I take as representing the opposition, bears on the adaptability and the correlated ambiguity of the thought experiment (which allows an '*intuition pump*'). I think that the ambiguity can be resolved by distinguishing three types of possible variations: **i)** the realizability of a certain assigment is considered without the constraint of time; **ii)** the constraint of time is applied and forces the parallelisation of certain operations; **iii)** the constraint of time is applied and prevents the possibility of the program to be executed. For type **iii)** nothing can be concluded, although it is reasonable to extrapolate from the two other ones. For type **i)**, it is clear that the program can be realised by a (serial) computing machine manipulating symbols in the set $\{0, 1\}$ on the model of the initial thought experiment and it is clear that I do not understand what the program is doing (its meaning) unless I actively research this, but this is not part of the thought experiment. This applies in particular to deep learning algorithms. For instance let us consider the algorithm which consists in training a neural network to distinguish pictures which have a '*smurglof*' in them and ones which do not and then apply it on a particular picture. If I execute this algorithm serially, considering the bits of information that the pictures contain one by one, I will never form, during the training, a concept of what patterns (the smurglofs) the pictures contain that actually distinguish them. The output of the algorithm informs me if it contains a smurglof or not, but I do not have any idea of what this *means* because the word does not refer to anything in my experience. For type **ii)**, the parallelisation of operations can lead to the construction of such references. For instance if I execute the above algorithm by looking at the pictures (not bit by bit), I will naturally form a concept of what a smurglof is and actually *understand* what the word means. In this case the thought experiment has a less straightforward effect because some effort is needed to see that the point may still be valid. The reason is that the construction of reference is there accidental and not necessarily related to the execution of the algorithm itself. What this means is that it is in principle possible to inhibit this construction and arrive at a similar conclusion as in the case of type **i)** arguments.

It seems that the changes operated on the initial argument, in order to exhibit what is essential to it and make its meaning clearer (J.Searle) or in order to prove that it is possible to arrive at different conclusions from the ones of J.Searle with similar arguments (D.Dennett), consist in distorting the argument to make it of type **i)** or **ii)**. I think that while the type **i)** arguments are valid, they do not cover all the possible situations: in particular the conclusions to be drawn from type **ii)** are unclear without a better understanding of understanding - and this leads us to further creative introspection.

In any case it is clear that the purpose of J.Searle is not to deny that minds are mechanical, only to doubt that artificial intelligence is useful for the mind-body problem: "*I begin with that old chestnut of a question: 'Could a machine think?' The answer is, obviously, yes. We are precisely such machines. 'Yes, but could an artifact, a man-made machine, think?'*" [Searle]. In fact it is in principle possible to describe the mind in terms of programs, however not programs which can be implemented in machines as we conceive them, but only describe causal relations between mental events. This mechanistical account would not exhaust experience (for mental events themselves are not all causal relations). In fact I like to treat causation as only a *layer* of the experience, other layers including particular types of phenomenal experiences (visual or auditive) as well as transcendental operations on phenomenal experience. I think that it is reasonable to approach consciousness through the spectrum of some of these layers, but not, like integrated information theory, to think that these layers exhaust experience. For instance constructing a correspondance between the structure of the human brain and the causal layer of experience based on isomorphism of structures may produce hypotheses on how the other layers may be related to the physical, or on the '*location*' of the point of view; this can enhanced by an exhaustion of what in the experience is structurally analogical to the brain before moving to parts which are not. I arrived at the idea that phenomenal experience exists in additional dimensions to the four dimensions of space-time - in fact if you think about it a little bit you may see that this statement is almost trivial, but I think that it can lead to some interesting change of ideas. In particular I believe that it is possible that the organisms to which we attribute conscious experience exhibit information integration in a way that can not be conceived to occur inside the four-dimensional space-time, for dimensionality constrains the possibilities for information processing and its organisation. Such an organism would need activity in additional dimensions (conscious experience) in order to integrate information the way it does. In this sense, an organism may have conscious experience only when passing a certain threshold of information integration, eliminating panpsychism. The quantity of information integration may characterize organisms having conscious experience (the ones which integrate information over a certain threshold), but not how this experience is related to the physical. One way to do that may be to search, provided a certain apparent integration in the four-dimensional space-time, how this integration can be realised with a minimal number of additional dimensions and ressources in these dimensions.

Digression: *I believe that one source of D.Dennett's illusionism lies in his article Where am I ? [D78], in which he proposed, through a narrative, a thought experiment which consists in imagining a surgical procedure allowing to separate the brain from the remainder of the body without altering phenomenal experience. In this context it is difficult to find an answer to the question 'Where am I ?'. Although D.Dennett develops his thought experiment further, in this situation both ideas that I am in my brain and that I am in my body are intenable. I think on the other hand that the confusion comes from the idea that 'I' am in the four-dimensional space-time, wether it is thought in the classical or relativity way, which occurs with the identification of the interior of the mental space with the interior of the skull.*

II.6 – H.Putnam

Along the approximation process that I mentioned in the last section, there are methodological temptations that I would like to keep away from. One of them, rooting H.Putnam's functionalism, is to approximate the actual mind using functional equivalence to identify mental states to functional states of the physical part of the organism, in order to circumvent introspection - an idea which probably roots itself the project of integrated information theory to relate mental states to parts of the *causal structure* of the brain). I will explain here why I think this is a mistake, reading through a pair of papers of H.Putnam.

II.6.1 – About the identity between different mental states and the functional hypothesis

In his article *The nature of mental states* [P67], H.Putnam defends a form of identification between physical states and mental states which allows to talk about mental states through physical states while ruling out a complete identity. For instance for H.Putnam, temperature (mental state) *is* mean molecular kinetic energy (physical state). Of course these two concepts are different because temperature refers to a sensation while mean molecular kinetic energy refers to a mathematical formula - in fact you may replace in this formula the energy with a hypothetical elementary sensation of energy, and this results in a transcendental formula not reduced to causal terms, but this is not the way it is usuall thought. In fact some arguments have been provided against this complete identification, such as the fact that it is possible for someone to experience a mental state without knowing that the correlated physical state happened. I find these amusing because the formulation of the sentence itself relies on the distinction between two different entities - a distinction for which one has to use introspection, without which the question of the relation between mental states and physical states makes no sense.

On the other hand they are identical in a sense that is close to the meaning of the equal sign (=) in mathematics. Let us for instance take the equality $3 \times 15 = 9 \times 5$. One can think about the two sides of the equality, 3×15 and 9×5, as classes of instances of mental processes. Take a mental object - for the example let's take a mathematical point. The first of these processes consists in considering a group of 15 points and '*copy-paste*' this group of points three times on a mental board. The result of this process is the number of points that you can see on the board. The second process is similar but takes a group of 5 points and copy-paste it 9 times. The processes are different but they are *equivalent* because they have the same result. You can also think of this in causal terms: if the result of the first process is the integer n then the result of the second process is n, and reciprocally; then if you consider embedding one of these two processes in a larger one, you can choose any of them for they are *functionally equivalent*: whatever is caused or causes one respectively is caused or causes the other. This is the meaning of the equality $3 \times 15 = 9 \times 5$: the two processes are equivalent related to their further use in mathematical constructions, and thus can be considered as completely identical *in this context.* Transposed to mental states and physical states: the temperature and mean molecular kinetic energy are not identical but are functionally equivalent and thus can be identified.

This is the conception of identity that H.Putnam uses in order to render conceivable the identification between different mental states (in particular strictly mental states and physical ones). With this type of identification in mind, he proposes that mental states are identical to functional states of the organism, meaning a local (in time) set of rules of conduct, in other words causal relations. In principle the interest of this identification is that causal relations can be computed on causal accounts of the organism, from a third-person point of view. In order to support this hypothesis, which I shall call the *functional hypothesis*, he rules out other proposals of identification of similar nature: **i)** the idea that mental states are physical states (of the brain); **ii)** the one that they are behavioral patterns.

The hypothesis **i)** is improbable, according to H.Putnam, for it is improbable that evolution could lead to only one physical correlate of a mental state, let alone for all mental states. This is the idea of *multiple realizability* (by physical states) of mental states: a physical state can cause a mental state, but this is not reciprocal - in particular it may not be the only cause of this mental state. On the other hand it is possible to prevent an animal to exhibit any reaction to the induction of pain, while this animal still feels pain: this proves that the hypothesis **ii)** is not acceptable. Along the same line of thought I could add to the argument of Putnam that one can also simulate the behavior of someone in pain (there lies also the threat of illusion: here we come back to the remarks on M.Grazziano's ideas). This means that there can not be any causal relation between a behavioral pattern and a mental content such as the feeling of pain.

These arguments do not affect the identification with functional states, for it is not in principle possible to simulate these functional states, and that functional states satisfy multiple realizability. However this hypothesis is convincing only because the means to refute it are lacking: how do we mesure a functional state, provided the fact that causation and furthermore functional equivalence are unstable by approximation ? - let us note that the situation here is different than the one of mathematics, for in this discipline the experiencial context is fixed beforehand. As a matter of fact, integrated information theory suffers from similar problems - I have argued in this direction in another article [G20]. More precisely the problem is that we do not have a *distance* which evaluates how far the functional structure of the approximation made of the organism considered is from the functional structure of the actual organism and thus how relevant the computations made on models of the organism actually are. I think that this problem arises when attempting to circumvent introspection, for this distance in question (although obscure) is provided by introspection itself.

I also believe that it is possible to recover the idea that mental states are related to functional states of the whole organism with the thought that mental states lie in additional dimensions, in the sense that they consist in causal intermediates between physical states (in particular executing the function of integration). In this setting, understanding the relation between the mental space and the physical space consists in having a faifthful representation of how they are spatially articulated - without this I believe there is no way to explain how a mental state and a functional state may be functionally equivalent. This articulation can a priori take many forms, but I think that an identification based on causal equivalence similar to the one used by Putnam between mental states which lie in different dimensions - such as what we usually call the mental and what we usually call the physical - can be only partial - where a systematisation of the identification would result in a form of physicalism - and through causal chain which go through the intersection of these dimensions (here I use the term dimension for the mathematical term of *vector space*, which, for mathematicians, should make more sense of the use of the other term *intersection*). In this point of view there can be still a experiencial (meaning in the time of experience) equivalence between some mental states and some functional ones, but this equivalence may be ultimately reducible to an equivalence between parts of them which are limits of elements of respective dimensions approached from these dimensions - and there can be a complete identity, not only experiencial but absolute. Furthermore the functional equivalent of a mental state could be necessary but not sufficient for the presence of the mental state.

II.6.2 – On the criticism of L.Shapiro

In order to be more precise on the position I adopt relatively to the thoughts of H.Putnam, I would like to spend some time on the article of L.Shapiro, *Multiple realizations* [S20] in order to separate in the criticism of functionalism the aspects I believe in and the others I do not. Ultimately the aim of this article seems to be a redirection in the definition of special sciences, meaning the sciences which are concerned not with physics themselves but with particular physical systems, or *kinds* of physical systems. According to L.Shapiro, special sciences should not find a priori laws of functional kinds - where the function defines the *kind* of systems studied - but focus on empirical grouping of systems of the same kind in order for a comparison between the particulars to shed light on the reasons for the form of the realizations of the function. In order to support this view, he questions the idea of multiple realizability as proposed by H.Putnam and then generalized by J.Fodor to defend from attempts to reduce special sciences to physics.

The text itself lets appear two grounds for the criticism of the notion of multiple realizability. The first one is related to the problem I mentioned above of the approximation or the distance between the representation - in particular functional - and the objects of experience - the organisms considered. There is a tendency, considering machines as we conceive them, to indentify the machine as a functional description of the object whose presence is experienced and the object itself (this tendency is similar to the one I mentioned in my other article between mathematical objects as concepts and how they are instanciated in the experience, a tendency enhanced by the fact that there is a one-to-one correspondance between them). As L.Shapiro puts it: "*But computers are not machine tables, nor are they intended to be machine tables. Rather, computers are devices for implementing the sequence of functional relations that a machine table describes.*" In fact for the machines that we build, this identification is natural because, in the circumstances of their use, the table describes accurately *all the possible behaviors* of the machine. Problems appear when using analogy between machines that we build and natural mechanisms, usually on the basis that they are both *mechanical* - however we should differenciate machines and mechanisms! The difference is that in the case of natural mechanism, the identification does not hold because the description is only partial: we do not have access to all the possible behaviors of the mechanism. Institutions for instance work mechanically, but they are not machines for the description of how an institution works concerns only a subset of all the situations of '*use*'. For instance the human beings which '*compose*' the institution have emotions which are neglected in the description of the purpose of the institution and under particular circumstances can affect its function in a way that is not predicted by the description. The situation is similar for natural mechanisms; in fact L.Shapiro uses the example of two computers executing the same operations in a particular sequence of events but have a different computational power: the difference is not seen in current situations of use but the difference may appear in some other *possible* situations. For L.Shapiro, "*Once we see the distinction between a description of a system and the system described, the temptation to move from claims of functional isomorphism to the truth of MRT [multiple realizability thesis] loses its allure.*" The reason is that if we do not have access to a complete description of a natural mechanism, how can one judge that two of these mechanisms are really different realizations of the same kind ? For instance it is not clear what would make an octopus eye and a human eye different realizations of the same function and not two different human eyes different realizations ? A similar argument is posited by L.Shapiro [about corkscrews and their colors], but my point here is to stress that the problem of multiple realizability here comes not from the thought of the natural as mechanical but of as descriptible with finite machines.

The second ground is the ambiguity of the notion of functional description: how should the function of a natural mechanism be described ? There are two possibilities to do so: else we only look at a function in terms of its input and outputs - the description is of what the mechanism does - or the description includes the particular causal relations which it consists in - here the description includes *how* the mechanism does it. A function is then multiply realizable when there are at least two different '*kinds*' of systems from which it is possible to extract this function. The term kind here is more intuitive than properly defined. It seems that H.Putnam chose the second type of functional description: "*As Putnam defines this concept, 'Two systems are functionally isomorphic if there is a correspondence between the states of one and the states of the other that preserves functional relations'*". The problem of the resulting notion of multiple realizability is that the equivalence between two systems is trivial: they can differ only by inessential (non causal) aspects of experience. In this case multiple realizability is philosophically non-informative. However it is possible to rectify the notion by considering the weaker version of functional description and that different kinds of realizations consist in different types of causal structures (they differ by the strong version of functional description). Although the notion of type here is not well defined, I consider that it is conceivable to find a precise definition. This way the possibility of non trivial multiple realizability is immediately verifiable, but this leaves open the problem of the possibility of multiple realizability for a collection of functions, and the philosophical conclusions are a bit weak for talking about the relation between the mind and the brain, but for my purpose it is enough to say that this kind of multiple realizability is in principle possible and this makes the thesis of identity between mental states and brain states as weak, and this leaves open the possibility to think about the relation between mind and brain otherwise.

II.6.3 – Another approximation problem: the τέλος

I would like, before going further, to insist on the multiplicity of approximation problems which prevent from talking positively about the mind through its principled identity - in the sense of H.Putnam - with the conceptualisation (in particular causal structure) of behavior. One other example of approximation problem is the one of the τέλος. For instance the function of an eye is to see, but the τέλος consists in the particular use of the eye, in other words what it is made for, which affects in particular the way it is connected to the remainder of the brain, positioned relatively to it, as well as the particular way it executes its function - in particular from points of view where the functional description does not exhaust the behavior of the system in question.

This is another point where the approximation process applied to behavior instead of mental mechanics via introspection fails to be informative about the mental, and where the argumentation of H.Putnam becomes fragile in a subtle way, for instance in his article *Minds and machines* [P60]. I use the term subtle for the reason that, in contradiction with what is usually thought, H.Putnam does not adhere to the thesis that machines can think or that human beings are machines - here in the sense that they can be constructed - as he explicitely states it in the end of the introduction of this article. In the text the main use of computing machines is to construct an analogy between the problem of the relation between mental states and physical states and the problem of the relation between *logical states* (the ones that are written in the machine table) and *structural states* (the physical states of a computing machine as physically realized) of a computing machine. With various instances of this analogy, he proves that it is possible for mental states and physical states to be different absolutely but identical in the sense described above in Section **II.6.1**. With all this I agree but the problem lies in the philosophical conclusions offered in the Conclusion: H.Putnam conflates the *mind-body problem*, which I believe is the problem of *how* the mental and the physical are related, to the problem of *how is it possible that the mental and the physical can be identified ?* While the series of analogies used in his article to show that the later is a purely verbal problem, they do not tell us anything about the mind-body problem. The reason is that while machines can realize functions related to the states it is in (wether logical or structural), like ascertaining that it is in a certain state, the machine never does it in the particular way a human being would execute it - for the actual realization of the function is shaped by the τέλος, approximated in the analogy.

Accepting these analogies, for H.Putnam, we should **(a)** accept that conclusions about the mind-body problem could not lead to shed any light on other problems "*of more than purely conceptual interest*" (like the question of wether or not human beings have 'souls'); **(b)** find a description of the mind as a machine; **(c)** accept that human beings as well as computing machines have souls. H.Putnam leaves the reader decide which alternative he or she wants to adopt, but clearly tends to choose the conclusion **(a)** and assumes that the reader who would not like to fall into absurd statements would agree with this conclusion. However it is derived from analogies which do not hold because of the approximation problem.

I thus leave open for myself the possibility that some progression on the mind-body problem *can* shed light on the nature of human 'soul' - what I may call the experiencing subject - and its structure. For this one has to accept not to force oneself to hold to a correspondance between mental content and (functional account of) behavior in order to inform about the mental, and allow the use of (creative) introspection.

II.7 – Global neuronal workspace theory

The global neuronal workspace theory, supported by S.Dehaene J.-P.Changeux, is the attempt of understanding the mind which I think is the closest to the approach I am drawing in the present text. However I believe that it is theoretically limited to experimental confirmation and precision of existing insights - about the spatiality and organisation of the mind - by the unacceptance of introspection as a method of study in a similar way as the approach of H.Putnam. I would like to argue for this in this section.

II.7.1 – Intuitive roots of the theory

I believe that the central idea of the theory is to find in the structure of the brain the structure of the mind drawn originally by B.Baars in his *Global workspace theory.* The rooting idea of B.Baars is that the human mind can be divided into two cognitive '*places*', one in which the subject can hold and transform a certain limited amount of information available *globally*, meaning along any of the possible cognitive actions (for instance moving an arm or focusing on a part of the visual space to collect an information in it and make it accessible globally) that the subject can enter into in the second one (which consists in the collection of the places which correspond to the cognitive actions), accessing complementarily to other information accessible only *locally* - in the analogy with a theater, the first place is the center, the scene, and the second one is the audience. Furthermore the first place is seen as central and the second as peripheral, and they thus consist in a spatialisation of the mind - conceptualizing the mental space as called for by J.Kim. Although it is a simplification of the structure of the mind, I take it as a rather convincing first approximation. The idea of S.Dehaene has been to build a correspondance between this division of the mind into two cognitive places and a division of the brain into two neuronal structures, and attempt to make more precise the cognitive traffic between the two cognitive places with the insight gained by the observation of exchanges of information between the two neuronal structures in question, as well as other intuitions gained by others with introspection. In particular in the first place cognitive operations - such as some the steps of a mathematical computation - are executed serially and consciously and in the second one they are executed massively in parralel and unconsciously - for instance the construction of a conceptualisation of the visual experience. I think also that the idea of D.Kahneman that the mind (as well as the brain) is divided into two coherent sets of processes, respectively working slowly and fastly in a complementary way - in order to reach an optimal trade-off between precision and rapidity - supports this vision.

The theory of S.Dehaene and J.-P.Changeux conceptualizes the division of the brain into two parts by associating particular types of mechanisms corresponding to the intuition: discrete (or digital) information processing for the center and analog and distributed information processing for the periphery; Turing's computing machines (with bounded memory) for the center and a collection of local processors for the periphery. In the theory, the local processors corresponds to part of the brain that have been attributed with a particular function (for instance processing specific senses information). Furthermore the central machine '*recruits*', '*connect*' and '*organise together*' the local processors, through a traffic of information through which some information available globally is transmitted to the local processors (their inputs) and some information held by them is made available globally (outputs) via a competition amongst many possible signals. The spatiality of the structure reflects the spatiality of division into two places of the mind, and each of the parts of the brain in question are considered to be neuronal networks (although in principle this could be adapted to more complex networks including glial cells for instance). The central network is called in the theory *global neuronal network* and gives its name to the theory. Another central correspondance is made also between the fact that a certain information becomes accessible consciously to the mind with the phenomenon of '*ignition*' of the global neuronal network, meaning that the access of the information to the conscious mind coincides with a global firing of this network. They also reiterated the idea of complementarity as a statement on the complementarity between two modes of information processing by neural mechanisms.

II.7.2 – Some inflexion in the introspective strategy

Despite the fact that the theory relies massively on intuitions gained via introspection, S.Dehaene seems to contort himself, in *The brain mechanisms of conscious access and introspection* [D13], with respect to how introspection should be used in the theorisation of consciousness. Not as a method, against philosophers and the '*ill-defined concepts*' that they posit in order to conceptualise consciousness as such [for instance '*pheomenal awareness*', or '*what it is like*'] (probably because as a scientist he underestimates the difficulty of the problem), because there are cases (of illusion) when the result of (straightforward) introspection does not correspond to the reality, but as an object (defending this use against residues of past behaviorism), because there are some cases of introspection which can be properly delineated, encountered similarly across subjects and thus studied experimentally, in particular when the subject has to report the presence of a simple stimulus (*conscious access*). Since what is reported via introspection does participate to the concept of consciousness, it should be possible to say something about this phenomenon via the experimental method.

This approach is reasonable, but as soon as this progress is made towards the conceptualisation of the object of study, reductionism gets over again. This abandon of the domain of experience happens in a different way than for integrated information theory though: here its reason is not the projection in an expected formal context but the restriction of the experiences considered to a thin subset of the possible ones. At a deeper level though, the '*understanding*' of consciousness that S.Dehaene claims often consists in unveiling brain mechanisms related to consciousness, but not in answering questions related to consciousness itself, such as how precisely these mechanisms are related mechanically to phenomenal experience ? For instance the following sentence ([D13],p. 4) is deceiving:

> "*We* ***understand*** *increasingly well* ***how*** *self-consciousness arises from a combination of brain circuits specializing in the representation of different aspects of our selves (sensory maps of the body, vestibular signals of head stability, programming of intentional movements, etc) (see e.g. Lenggenhager, Tadi, Metzinger, Blanke, 2007).*".

In fact by the identification of the mechanisms in question and their correlation with conscious processes, one may only have an information on what in the brain may be involved in these conscious processes, but not how these mechanisms are related to these conscious processes, let alone understanding this hypothetical connection. This kind of identification and correlation is what a strict experimental method restricts itself to by the rejection of introspection as a method - it is sufficient for the practical goal of the theory to provide a hint for the presence of an experiencing subject, but not for *understanding* it.

In the case of conscious access, the mechanism in question is the *ignition* of the global neuronal workspace and the global availability of information that is made possible by ignition "***is*** *what we subjectively experience as a conscious state*" (p.11). This identity, although it is attractive, may be the very result of the restricted context of the experimental study: in general - when considering any possible experience - the situation may be completely different. In particular the phenomenon of ignition may simply be a causal intermediate which happens to be forced in the restricted context. I think in fact that the global availability in the brain of an information could at most be identified to *global availability in the mind* of this information, which is different, until otherwise proved, from the mental state which corresponds to the information.

Furthermore if it is possible to accept this kind of identity, it is not necessarily the case that every aspect of conscious experience is 'identical' to a physical counterpart. Holding the idea that mental states may lie in additional dimensions, I think that activity in these dimensions occurs only when necessary (a sort of optimisation principle). The approach that I like reflects this principle in the sense that I have recourse to both identification between mental and physical and to additional dimensions only when possible and necessary.

II.7.3 – Towards a better understanding of the structure

Beyond the formulation of the identity between global availability of an information and the conscious mental state corresponding to this information, the theory accounts for other aspects of conscious access than the simple presence of the mental state to the mind:

> "*The GNW theory accounts for at least three aspects of subjective experience: (1) individuality: the same stimulus may or may not lead to conscious ignition, and whether such ignition occurs, in a given brain, is a stochastic event unique to each individual; (2) durability: thanks to its reverberating self-connectivity, the GNW network can maintain information "in mind" for an arbitrary duration, long after the actual sensory stimulation has vanished; (3) autonomy: the shaping of spontaneous activity by GNW circuits leads to the stochastic endogenous generation of a series of activation patterns, potentially accounted for the never-ending "stream of consciousness".*"

Other accounts of this type appear in other texts [for instance, differences between minds and machines, such as the difference in effectiveness in executing simple arithmetic operations or face recognition], however they most of the time consist in a simple reformulation, based on vocabulary of the theory, of a description which results from introspection around a conscious access event. Moreover: none of them concern aspects which are not straightforwardly connected to conscious access; all of them consist in pointing at the possibility to realize on the model aspects which are analogical to ones of the mind rather than constraining how these realizations should be done to be coherent altogether.

For instance it would be interesting to answer the following questions: **(i)** does the phenomenon of *ignition* concern only simple informations or also more complex ones ? which leads to the question **(ii)** should the ignition phenomenon be considered to correspond the propagation of an information through the whole global neuronal workspace, or as a warming of the neurons of this workspace in preparation for wiring this information from one local module to another (allowing a controled information integration) ? **(iii)** In both cases, what binds the quantity of information which is transmitted through the workspace ? this kind of considerations can provide a hint for the question **(iv)** how is an information retained by the workspace ?

In order to understand better the architecture drawn by the global neuronal workspace theory and its relation to the structure of the mind, I think it would be more significant to account, for instance, for how fundamental transcendental operations (such as the ones of integrated information theory) could be coincide (partially) with physical processes. In order to do so, one should reverse the approach in the same way as integrated information theory does: going from introspection and the description of experience to physics. Here this means describing experience in terms of elementary concepts and then making hypotheses on how this description can be seen or complement the model of brain already formed. Some of the aspects of experience that the theory does not take into account are for instance: **(i)** the location of the point of view (the *I*), and the self-attribution of its decisions: taking mental computing machines as a fundamental mode of the experiencing subject, how does it effectively decide if an action results from itself of a local module ? **(ii)** if the global neuronal workspace is used to interconnect local modules in order to instanciate locally in time a particular algorithm, how does the experiencing subject (or the machine) '*knows*' if this effort is necessary ? In other words how does it know if the algorithm in question is not already '*encoded*' in the brain ? **(iii)** how does it recognizes local modules when connecting them ? In other words what kind of '*adressing system*' makes this recognition possible ? **(iv)** How does the experiencing subject *makes present* to itself an object it holds memory of

? This object can be a number or a section of the visual space for instance. Does this presence mean that this information is held by the local memory of the global neuronal workspace ? In what sense ? Does the corresponding mechanism and the notion itself of '*presence*' differs from a type of objects to another ?

Before going further..

For the last question, we can observe in the practice of mathematics that drawings and notations are used to *make present* abstract objects to the mind. This operation has an effect on the possibility for the experiencing subject to conceptualise this objects and connect this conceptualisation to other objects in memory. In particular the *visualisation* of these objects make this more efficient (and easier). How is that so ? The practice of mathematics offers several more questions of this kind which may be used as constraints on the model: this leads us to phenomenology of mathematics.

Along this part I have defined an introspective approach of the structure of the experiencing subject, delimited its purpose and the way it operates to create meaning, more negatively (determining it by what it is not) than positively. The arguments used may lead to the thought that a science of consciousness is, despite recent development, not possible. However I do think that it is not the case: only its cultural weight, which do not participate to its essence, is responsible for the difficulties encountered. After these negative arguments I wish to provide more positive ones, taking examples of possibles directions for this approach to be applied.

We have seen from this part that we should shift from a theorisation of the mind as a machine to a mechanistical one. This change in the nature of theorisation signifies a change of purpose: from reproducing the mind (and in particular consciousness), a purpose which derives from the faith in artificial intelligence, to the possibility to recognize and act upon the mind. This can be made possible by condensating our designations of the mind used in the daily reality into statical ones.

III. – *How* may the human brain simulate mental computing machines?

This part is devoted to the following question: when simulating mentally a computing machine, what may correspond in the brain to the various elements in the definition of the machine - for instance the tape, the head, and its movement - in the brain ?

III.1 – Jacques Hadamard

In order to define positively the way I will use introspection for this question, I will take as reference the book *The psychology of invention in the mathematical field* [H45] by Jacques Hadamard.

III.1.1 – Criticism of H.Poincaré's psychology of mathematics - variations and constancy of the mathematician mind

I believe that the main point of the book is to offer a criticism of H.Poincaré's psychology of mathematics. In particular he criticizes the idea of H.Poincaré that the way a mathematician creates meaning is *determined* by the way this mathematician thinks [for him, as an analyst or a logician], and in particular independant from the mathematical objects considered:

"*The method is not imposed by the matter treated. Though one often says of the first that they are analysts and calls the others geometers, that does not prevent the one sort from remaining analysts even when they work at geometry, while the others are still geometers even when they occupy themselves with pure analysis. It is the very nature of their mind which makes them logicians or intuitionalists, and they cannot lay it aside when they approach a new subject.*" - **H.Poincaré**.

J.Hadamard opposes and explores the idea that mathematicians differ in the way they tend to use their mind in to create meaning rather than in the '*form*' of their mind. Some rely more on the unconscious part of their mind [intuition], while others do rely more on the conscious part [logics]. They are all situated in a spectrum in between these two polar behaviors, according to the depth the reflection is immersed in the unconscious part of the mind, and how connected their ideas and interests appear connected to others [dispersed for a more intuitive mind, concentrated for a more logical one].

In fact I think that paradigm shifts such as the one which happend during the XIXth century from constructive mathematics and conceptual ones also illustrate that the way mathematicians use their mind does not follow strictly intemporal pathways.

Digression: *I already mentioned the celebrated reflection of H.Poincaré of how the 'illumination' phenomenon through which the solution of a problem appears all of a sudden after a long unconscious processing after ingesting the data of the problem and failed attempts to solve it. For this also there is some criticism to be made: in some particular fields where mathematics are involved (I am thinking of statistical physics as I know it), the solution of a problem may happen to appear after a long process combining various techniques without this apparition being all sudden and clear.*

As J.Hadamard puts it, the way a solution to a problem is constructed consists in the position of many gatherings and display of mathematical objects and each time testing the adequation of this construction as a solution to the problem:

"*Indeed, it is obvious that invention or discovery, be it in mathematics or anywhere else, takes place by combining ideas. Now, there is an extremely great number of such combinations, most of which are devoid of interest, while, on the contrary, very few of them can be fruitful. Which ones does our mind - I mean our conscious mind - perceive?*" - **J.Hadamard**.

These combinations are mostly produced randomly (it makes sense to the observation that this processing can happen during sleep), and concern objects which are posited by relations made between objects which appear in the *conceptualisation* of the cognitive situation which is opened by the reflection on the problem with abstract objects which are not situated. Most of these operations - conceptualisation, extraction and connection between objects and their combination - happen mostly unconsciously. In fact it may be interesting to make hypotheses on how the brain may mecanically execute these operations sequencially (in particular in which order does it make the various possible combinations ?) When considering anew after this unconscious processing the cognitive situation, if a solution has been found it appears for it has been selected and stabilised unconsciously amongst other possible constructions.

Mathematician minds differ in the distribution of degree of consciousness for these operations and these variations root the spectrum mentioned above, however surely not in the kind of transcendental operations through which they create meaning. I believe that it is important to understand - and I have begun in my last article with a partial classification - the set of all possible ways for this meaning creation, through these operation, in order to keep alive all these possible ways.

III.1.2 – An example of phenomenological description for mental imagery

Besides these considerations, J.Hadamard strives to differenciate the practice of mathematics itself from the mathematical script. Quoting Schopenhaur - "*Thoughts die the moment they are embodied by words.*" -, he argues that when thinking, words are absent, and thus are not necessary to thoughts. I only partially agree with this for it depends on the way we think about words: thoughts die when they are embodied by words *because* we think about words as embodiement of thoughts instead of a way to communicate *a direction of thoughts*. However I believe that the purpose of J.Hadamard here is to point at the fact that mathematical words do not capture the reality of the mind of the mathematician, in particular they do not convey the *mental imagery* used by the mathematician in order to have an insight into the machinery of the concepts that are manipulated - in other words what is present to the mind in the moment of reasoning.

In order to illustrate what he refers to, he proposed a phenomenal description of what he has in mind consciously when proving that *there is a prime greater than 11*, describing the mental images that he holds at each step of the proof. This description is as follows:

STEPS IN THE PROOF	**MY MENTAL PICTURES**
1. *I consider all primes from 2 to 11, say 2, 3, 5, 7, 11.*	**1.** *I see a confused mass.*
2. *I form their product N being a rather large* $2 * 3 * 5 * 7 * 11 = N$.	**2.** *N being a rather large number, I imagine a Point rather remote from the confused mass.*
3. *I increase that product by 1, say N plus 1.*	**3.** *I see a second point a little beyond the first.*
4. *That number, if not a prime, must admit of a prime divisor, which is the required number.*	**4.** *I see a place somewhere between the confused mass and the first point.*

I mention this here for it provides an example of possible phenomenological description - and thus a direction for introspection - of mental imagery from which we may infer on the structure of the experiencing subject. We may notice that at each step of the phenomenological description differs from its counterpart in the proof by its simplicity, since the objects of this description are only visual ones displayed on a '*mental board*'. In particular the '*confused mass*' and the points do not contain any arithmetic information that numbers hold as mental objects. I think that during the mental process which underlies the proof, what appears to the mind reflects the creation of connections between spots on the mental board and *local processors* (the same ones as in global neuronal workspace theory) which hold this kind of arithmetic information, thus organizing the communication between these local processors through the mental board. Once the connections are created, this communication happens mainly at the unconscious level. In order to refine this description, one may ask further questions such as *where* is the mental board and how connections are realized.

While purely subjective, mental imagery does not seem to be dependant upon the particular subject - this should be clear for mathematicians and, according to J.Hadamard, for some intellectual from other fields who use similar mental imagery for their reasoning. Descriptions such as the one above, if properly executed intersubjectively, and when compared to the model of mental space structure, shoudl thus tell us something about the structure of experiencing subject, and not be specific of any particular subject. Such a description may be done on other mathematical reasonings or other aspects of mathematical practice, and as well simple daily cognitive events. For instance what appears to the mind while searching for a memory which I know I have but is lost ? (this provides an interesting example of simple dynamics between unconscious mind and conscious one, for it consists in a partially conscious process directed towards the unconscious mind so that something appears to the conscious mind). However when focusing on mathematical practice, phenomenological descriptions obtained out of it should be more likely to be interrelated. Based on this interrelation we may lead abstract queries such as: what kind of mental images can I hold in my mind (characterization) ?

As I mentioned earlier, I would like to focus on computing machines because they form a fundamental mode of the experiencing subject. What I would like to do is introspecting in a similar way as J.Hadamard on: what appears to the mind when simulating a computing machine ? Furthermore what is the cognitive process by which I arrive at the conclusion that I may simulate the computations of any computing machine ?

III.2 – What can be hoped for - meaning creation

Before answering these questions, I will use one more section to explain the epistemological status of the elements obtained by the kind of reasoning operated here, in particular how they should be considered and articulated, as well as what may be expected from these operations. All this shall be explained by the properties of what I call '*iconic introspective capacity*'.

III.2.1 – Iconic introspective capacity

In 1960, G.Sperling investigated [S60] visual short term memory in experiments during which subjects were presented with a brief visual stimulus which consisted in an array of alphanumeric characters. After the stimulus disappeared they were requested: **(i)** to report as many characters from the stimulus and their position in the array; **(ii)** or to report as many characters in a particular row of this array (this row was not known before the experiment and this information was transmitted at to the subject under the form of an auditory stimulus). The experiments demonstrated that the capacity to report was significantly increased from the first case to the second one. G.Sperling coined the term '*iconic memory*' in order to designate the memory of visual displays. I believe that the difference in terms of capacity of iconic memory when focusing on a restricted area of a visual display may be explained in the following way: first, the information contained in the stimulus is held only for a certain period of time after which it is lost; second, the action of reporting requires a choice of order to follow on the positions of symbols to be reported, and the larger the area of focus, the longer the period of time required for choosing; third, the longer the time required for choosing, the less accessible the information is.

I think that this paradigm can be applied to the capacity for introspection, which can be thought to consist in the report to oneself of elements in spatial displays, where the geometry of the space varies from one area of cognition to another. In particular the capacity to introspect in one area is dependant upon how wide it is. In terms of [G20], statical designations and dynamical ones represent two complementary moment of the language which describes the world: the second ones, when composed, delineate areas of cognition in which introspection can act in order to constitute statical designations which may be selected and composed and altogether structured in order to *understand* these same areas. Furthermore the notion of micro-world statically designate examples of this notion of area of cognition.

III.2.2 – How to apply this paradigm ?

Practically, however, how should this paradigm be applied to introspection without an understanding of the spatiality of the mind ? One way is introspect while holding a certain object in mind, wether it is a singular information, a concept or a question - for instance, how does the adressing system of my conscious mind works ? This objects works then as a selector which restricts the subarea I am considering without actually characterizing it in the area I am introspecting in.

In particular, this is one purpose of focusing on a fundamental mode of the experiencing subject in order to understand its general form and dynamics, as well as focusing on the layer of causal relations between mental events rather than the phenomenal content of these events. Furtheremore some concepts such as free will - in particular facts related to these concepts such as judgements - may locally play a similar role, under the form of questions such as what information is accessible to 'me', at which point, how difficult is its access, how much do 'I' contribute causally to an event, how do I decide to attribute to myself the cause of an event ? etc. I will discuss these in later sections.

When considering mental computing machines and their simulation, I can hold as a question the form and localisation of the machine head, as well as by which mechanisms it reads writes and move. Is it moving or fixed (I think that it is actually partially fixed, partially moving) ? Furthermore how do 'I' have access to informations about the computing process ? Can an answer to this question explain why I have access to some information and not to another ?

Although they are of different nature, I would like to mention that other constraints coming from our intuition on space, time, and how these constrain information flow - for instance that there can not be an infinite amount of information in an '*element*' of space and time, as well as the partial knowledge of the brain architecture. This kind of constraints may be used to select between alternative hypotheses.

The elements of a conceptualisation out of introspection may be used further to increase introspective capacity; any time it is possible it should be assumed, at least temporarily, that these elements have a material counterpart and consider how this counterpart may be related - both spatially and functionally - to the elements of the model already formed.

As well the use of introspection could make reflexive and one could attempt introspectively to explain the properties of iconic introspective capacity itself.

III.2.3 – Conceptual creation and mathematics of the experiencing subject

I believe that most barriers on introspection are related to this notion of iconic introspective capacity, including the dynamical nature of the experiencial content of a considered area of cognition. In point of fact I have believed for a long time that the nature of the experiencial content determines the possibility of a reliable conceptualisation. Thinking this way implies that the creation of concepts may only come with an increase of complexity. On the other hand the examples of A.Turing's computing machines, the general definition of dynamical systems or even J.Nash's game theory made me wonder: how come that these 'simple' mathematical objects were not defined before ? With time I came to the conclusion that an meaningly effective mathematical concept is not contained in its definition. Regardless of the complexity of its definition, its experiencial roots, which lie in between the intelligible and the sensible, are determining the questions with which mathematicians inquire it and ultimately understand it. A significant concept - such as, I believe, the one of causal structure - is so for the depth of these roots. I think that similar concepts may come out of a clear picture about the structure of the relation between the mental and the physical in the case of mental computing machines.

Simple but significant concepts may be drowned under the sea of all possible concepts until they are defined. As J.Hadamard puts it:

"*Invention is discernment, choice.*" - **J.Hadamard**.

III.3 – On the simulation of mental computing machines

In this short section, I will dwelve into the main subject of this part: the simulation of mental computing machines. Its purpose is only to provide evidence for the possibility of further developments in the same direction.

Let me remind that a computing machine, as defined by A.Turing, consists in an infinite tape written with symbols in a finite set of possible ones, on which a machine head can move, read and overwrite symbols. The machine head can be only in finitely many possible states and a finite table describes the dynamical behavior of this head, determining its movement and what it writes according to its current states and the information on the tape it has access to (which is written on the tape at its position).

As I mentioned above, this concept was formed to describe the mathematician's mind in the process of computing. In fact once properly defined, this structure may be recognized in many other mental processes. Furthermore it is relatively intuitive that the human mind may simulate any computing machine. The question I would like to address here is the following: *how to describe what is happening in the mind when I make a judgement about a statement of this kind* ? It is quite clear that *what is happening* comprises the (abstract) computing process itself but as well can not be reduced to this computing process. In particular, what may correspond to the elements of the definition - tape, machine head - in the brain ? Are there '*auxiliary*' processes which are necessary to the simulation ?

I present my observations in a table similar to the one of J.Hadamard, where the proof is replaced with the mental simulation of computing machines, and mental pictures with mental processes which underly each step of this simulation: see Table 1, Table 2 and Table 3. I also assume that the machines are initialized with empty tape.

STEPS IN THE SIMULATION	**(TRACE OF) MENTAL PROCESSES**
1. *I display the table which is meant to contain the machine rules.*	***1.*** *I choose a random place of my 'mental board'; I mentally instanciate a table at this place; only one or two cases appear to my mind with the algorithm of repeating this display according to the quantity of information it will contain;*
2. *I display information relative to the rules in the table.*	***2.*** *I determine the number of rows of the table (5) ; I connect each row of the table with the idea of a type of information that it will contain: current state, current tape symbol, new state, new tape symbol, movement direction*

Table 1: Phenomenological description, part **i**.

This is a description of '*what happens*' in my mind when I imagine simulating a computing machine, whatever are its rules and symbols. Based on these mental constructions I can confidently say that I can simulate any computing machine - I only need material and symbolic inputs. Now let us try to see how each of the operations involved in this process may coincide with physical processes in the brain.

STEPS IN THE SIMULATION	**(TRACE OF) MENTAL PROCESSES**
3. I determine how to use the table.	*3. for the remainder of the simulation, I hold the information that at the beginning of the computing process, I will receive information (for instance from another person) of the machine rules and write them down in the table (one column, one rule); I split the rows into two sets (inputs,outputs); during the computing process, each time I access the table, I access it with two informations which determine a column to look into, these two information corresponding to the inputs; while looking into the column, I drop these information from my memory and collect the information contained in the last three rows of this column, before leaving the table.*
4. I display the tape and the machine head.	*4. As for the table, I choose another random place of the mental board far enough from the table (so that they do not overlap: this simplifies information processing); I mentally instanciate a tape at this place, meaning a table with only one row (of undefinite length), hold the information that the machine head will be on the leftmost position at the beginning of the process in a particular state. I picture that in general the machine can be on any position (I see one point on a line) in any state (I see another point on the top of the first one).*

Table 2: Phenomenological description, part **ii**.

Let us begin with the instanciation of the table and the tape. I think it is reasonable to think that the mental board that supports them coincides with neural maps of grid cells networks, and that the instanciation of these objects consists physically into 'warming up' areas of these networks by simply 'visiting' them. I think that it is also reasonable to think that the information of the algorithms allowing the construction of these objects - when actually executing the process - is not 'contained' in this network but in another specific area of the brain and that they are hard-encoded into neural structures. I see two reasons for that: first, I imagine it would be a difficult process to script the algorithm from scratch and this difficulty does not match the rapidity with which I instanciate objects in space; second, it is likely that I use this kind of elementary algorithms for many other situations in which I am not necessarily conscious that

STEPS IN THE SIMULATION	**(TRACE OF) MENTAL PROCESSES**
5. *I determine how to use the tape.*	***5.*** *When entering the area of the tape holding a certain information from the board, I use the first one to overwrite on the letter, the second to change the state of the machine head (I imagine respectively the act of erasing the content of the points and then droping the the information I am holding), and the third to move the machine head to a neighbor position (I imagine erasing the two points and moving both points to the left, erasing them and writing them back to the previous position, then erasing them again and writing them on the position on the right), before leaving the area.*
6. *I determine the visit order of the table and the tape.*	***6.*** *I hold the information that I will begin the process in the area of the tape; I also picture a movement of back and forth between the tape and the table.*

Table 3: Phenomenological description, part **iii**.

I use them, and it is simpler to script them in a dedicated area of the brain and 'call' them whenever needed. When I instanciate these objects (the table and the tape), I imagine that I create (locally in time) a connection between the location of the algorithms and the areas of the network. Here again I think it would be difficult to make a 'direct' connection, meaning connecting specifically the algorithm with the area: if the connection was not there, my brain would need to create a neural path that it would need to destruct imediately after; if it was there, there would be no reason for the many other potential connections of this type to be already there in the brain, however there would be a problem with the finiteness of information present in an element of space and time : how could all these neurons be present in the brain at the same time ? I think on the other hand that it is reasonable to believe in a hierarchy of hard-wired access connection from the global neuronal workspace to the many areas of the brain. The connection between the algorithm and the area of the grid-cells networks may be thus created by sending a signal from the global neuronal worskpace to the area of the brain containing the algorithm, back the the global neuronal workspace where the signal is directed to the area of the grid-cells networks which has been warded up at the same time, creating a path between the two - this path may be maintained all along the computation process, simply by continuously sending a similar signal, which is then directed only by its own trace.

Since the apprehension of 'I' does not depend on time during this process, in particular the fact that I am constantly able to designate 'I' as the cause of each sub-process, this indicates that the place which may be the support of apprehension processes is also constant, somewhere in the global neuronal workspace. When I say that I visit a certain area of the mental board, or the area containing a certain algorithm, I am present in this area, but I also still am where 'I' is. One way to see that this is possible would be to say that I consits in a signal continuously sent from the place of 'I' to the cognitive place that I am considering, with immediate return to the place of 'I'. For instance I have hypothesized above that when I am creating a connection between an algorithm and a place, I am visiting these two places: this means that there is a signal sent from the place of 'I' to this place and immediately in return to the place of 'I', and the direction to which this signal is directed is continuously informed by information held in the place of 'I'. How is this signal directed ? Regarding specific algorithms, it is possible that the information of the adress of this algorithm is simply encompassed in the concept of computing machine as it is learnt (and then in the concept of table or tape, etc), this address being passed continuously when direction is needed to the place of 'I' in order to inform the signal's direction. When the adress is only partially defined, such as 'a place on the mental board' the direction of the signal is feeded only with the adress corresponding to the board itself, and then not feeded anymore with information, which renders the place in the board random.

Let me just remark at this point that the signal sent from the place of 'I' to the place visited at a certain time and back to the place of 'I' together with this place of I may be seen as the head of a machine working on a tape which has a more complex geometry than the unidimensional tape of the definition of computing machines. Furthermore the distinction between the tape and the machine head in this case is less clear.

We can consider then what happens when I am visiting the area of the table. In particular during this visit I collect some information. This means marking the sub-area where these informations are contained and connecting this sub-area with a counter algorithm, each value of the counter corresponding to a row of the table. In fact when I am moving from one area to the other, I am not really holding in mind these precise informations, which indicates that I have dropped them somewhere I will be able to gather them back when I will be elsewhere. Since there is a local route (via the global neuronal workspace) from the area of the tape to the area of the table, I only have to send a request to the information I need via this route, the address of the information in the table area being contained in this area, and get back the information at this address. Each time sending this request, the counter algorithm increments and sends an ending signal when the counter has maximal value. A similar process is exectuted when I enter the table with an information relative to the tape.

When moving from one area to the other, the 'I' signal only has to follow the route traced between the two areas. In order to complete the picture I only need to account for what happens when writing on the tape. When I have instanciated the program that I am following, I have only connected the area with an algorithm including the action of modifying certain symbols contained in the tape. However when actually executing the program materially (for instance writing on a sheet of paper), it is not difficult to figure out by what means I will execute this part. It is possible that the prefrontal cortex contains intrasensioral commands which are applicable whatever the situation, and in particular here unconsciously, such as '*whenever I want to make a modification in my visual field, I have to use my hand*'.

Some remarks: (i) Let me remark here that when preparing the simulation of a computing machine, the actions of 'warming up' the various areas and the creation of local routes between them may be done in a different order than the one I presented. However this should not change what can be hypothesized about the relation between the mental and the physical. **(ii)** The second remark I would like to make is that when preparing my mind to execute a program like a computing machine, in the interpretation that I provide, 'I' mainly interconnect various algorithmic modules through the global neuronal workspace. Furthermore the kind of information that 'I' keep in mind across different tasks seem to be only the routes connecting the different modules. The phrase '*I keep in mind that the in the table in column x and row y there is..*' is a language construction: I do not really keep this information in mind, I only keep a way to access it without effort. **(iii)** The above may be seen as a collection of observations which can serve to sharpen introspection around this fundamental mode of the experiencing subject. In fact we may recall the question wether the mind works as a computing machine or not. There are elements to think that it is the case, but the above does not really provide an answer. However if we are to think that it is not the case, further exploration may lead to exhibit mind mechanisms which can not be described in terms of computing machines. **(v)** The approach that I present here does not pretend to be coherent with the data we have on the human brain, however the main point is not to provide evidence for a certain interpretation of data, but to make possible the collection of 'data' on the mind.

III.4 – Directions for further investigation

The purpose of the last section was only to sketch an introspective approach of the simulation of mental computing machines by the human brain *in practice*. In the following I provide some orientations for further investigation of this type. By reason of the nature of this investigation, they take in general the form of *constraints* which may ultimately determine our concept of how the mental computing machines - and ultimately mental machinery in general - are actually simulated by the human brain, by removing possibilities that we may imagine (when they can be formulated in terms of the model already constructed) and which enter in contradiction with the constraints, or by tightening the meshes of introspection.

Because the focus here is on the general structure of the experiencing subject, constraints may not be necessarily directly related to the concept of timeless experience and its fundamental characteristics, or fundamental operations acting on the representation of an experience, such as in *Integrated information theory*, but shall be concerned with experiencing, and thus the dynamics of the relation between the subject and its *Experience*.

In this setting I see two types of constraints: **(i)** principled constraints on the form of the experiencing subject; **(ii)** existence of particular modes of the experiencing subject and their characteristics.

In general constraints of type (**i**) apply on the constructed representation of the structure of the experiencing subject through an optimisation principle. Let us consider for instance the phenomenon of information integration. We may study this concept statically by considering particular experiences and wonder how to relate it to objectness. In the present context we should consider it dynamically instead, and how information integration is modulated for a certain purpose. If we query the question of why information integration *happens* or not while keeping in mind that its modulation requires an effort, we should arrive at the idea that it is optimised for the purpose, constraining the dynamics of the experiencing subject. In particular the idea of a central area where certain information is shared between many parts of the brain in a single time could in fact be derived from the optimisation principle, as it is less costly than sharing this information through a collection of module-specific connections. This applies also to the idea of a balance between permanent and temporary channels for sharing information. In practice this kind of constraint may be applied on the current model by wondering if there could be another system which is more optimal way regarding these constraints - such a system may account unexpectedly for other aspects of experiencing. On the other hand, constraints of type (**ii**) apply on the model of the experiencing subject by queries on the possibility to adapt it in order to account for some characteristics of the modes of the experiencing subject in question.

In the following, I will list and sometimes discuss some examples of such constraints.

III.4.1 – Specific modes of the experiencing subject

I have mentioned at the beginning of this section the functioning of the mathematician's mind as a particular mode of the experiencing subject. Some works, in particular by S.Dehaene (see for instance the presentation '*A close look at the mathematician's brain?*' [S17], or the article [SA16]), attempts to localize in the brain areas underlying simple mathematical operations, using experimental methods. I believe that in this direction the introspective method (that I have attempted to define above) may be applied on more complex operations and reasonings. I suspect this would lead to more precise, although hypothetical, insights in how brain processes may coincide with these operations and reasoning. However I have not developped significantly this direction yet, leaving it open to a potential reader (which may be the author).

However there are some questions that I find interesting to query: **(i)** when searching for the solution of a problem (or even more generally searching for a memory to remember), how does the brain underly the mental process of selecting and rejecting ideas or mental items based on the adequation to this search ? **(ii)** I have observed during my own training that the 'world' (defined as a local set of possible experiences) in which the subject evolves affects the capacity to conceptualize a particular experience and memorize and operate on this conceptualization. Is there a systematic relation here and how could this be translated into brain mechanisms ? **(iii)** On the combinatorics of instanciations of mental algorithms. I have analysed above the instanciation of a mental computing machine, which requires a certain mental effort: when instanciating a second one after the first, am I able to keep '*in memory*' the first one to repeat it after the second without the effort of re-instanciating it ? - if yes, how ? Similar question for the modification of an algorithm after modification of its formal description (these questions may be asked also for embodied mental processes). How do the answers to these questions change with the number of algorithms ? Furthermore how to characterize the effort demanded by the instanciation of a mental algorithm in its formal description ? **(iv)** why exactly makes it easier to process information when it is supported by visual items than when it is not ? For instance when I attempt to evaluate purely mentally some properties of triangles in general - let's say the equality of the lengths of all edges implies the equality of all angles - I have to construct in my mind the triangle point by point, and then construct a representation of the edges and angles, and come back to these objects to stabilize their presence in my mind. In can thus say that this construction is an obstacle for this mental reasoning compared to reasoning with visual support, but it is not all. In fact we can also put it in the following way: how to characterize what is purely mentally 'computable' ? **(v)** Some mental processes can be done in parallel: how to characterize sets of mental processes which can be parellelized ? what explains can the possibility of parallelization ? **(vi)** How to characterize properties of the mental space which determine the variable difficulty of the mental machine head (in other words the focus of my attention) to navigate in this space ? (could this explain why bidimensional grid structures underly cognition of space as we commonly conceive it ?). **(vii)** We may notice that when reasoning with visual support, we tend to subdivide hierarchically this visual support in a way which depends on the objects and operations done on these objects. Is there a way to explain or predict this subdivision from the data of the objects and operations in terms of actual computational capacity derived from the structure of the experiencing subject ? **(viii)** One may also analyse simple algorithms found in mathematical practice everywhere, such as visual search for an item or a type of item: what could they teach us about how mathematical objects may be '*stored*' in the brain ? For instance when, in a picture which consists in a white canvas with black dots and a unique red dot, I search for the position of the red dot. It is manifest that the search process is not accounted for by an exhaustive search over the canvas through a restricted scope until the scope's content matches the object searched for. In fact we could think in the following way: I can decompose the mental search algorithm, without affecting the way it is related to physical processes, into one step which consists in preparing the search algorithm to be executed without visual stimulus, and then another executing it on the particular stimulus. Let us assume I was asked by another person to execute this search: "*find the red dot in the picture and point at it with your finger*".

The preparation goes the following way: once parsed the sentence, I identify the imperative verb and the action it requires me to do, which points to an object in the sentence. I extract then the information relative to the object: its shape (dot) and its color (red). In order to do this, I need to mentally repeat the sentence by sending a signal to the module related to autitory stimuli while invoking a parsing algorithm and a semantic analysis on the sentence I heard, this by creating a connection between the module and the specific algorithm. Once the words relative to the objects are extracted from the sentence, I construct a mental representation of this object that I will use for the search, by invoking the construction algorithm for a point (which roughly consists in choosing randomly an area of the mental board and shrinking it until not possible anymore) and connecting it to the '*entry*' of the mental board, together with the concept of color, specified with the color red. Without contact with the visual experience, unconsciously (since the dot appears to me without effort), the signal bearing the information of the color red should be diffused through all the mental board, in a hierarchical manner - I assume here that there are along the grids hierarchical structures which encode positions on the mental board. The signal is compared on each position by an elementary process which results in a binary answer to the query. When positive, the answer is sent back to the entry of the board. In this case the signal which is sent to this position and sent back is maintained continuously until the end of the search process (this way adresses of the red dots on the board are encoded in these signals). These signals in fact may be the material counterpart of the conceptualisation of the visual experience, dependant upon a certain purpose - in other words they may coincide with mental objects. Once this is set up, 'I' only need to send a signal to the visual board holding in mind the idea that it will be directed in the board towards '*whatever has been detected*', following randomly one of the the continuously maintained signals in the hierarchical structure over the board. This description seems to account for properties of the mental search algorithm - in particular what is manifested in my mind when directing my attention to a red dot already present to my mind. More generally, what kind of characteristics make a mental object effortlessly searchable, and how do they affect the search process ? **(ix)** Many situations in the practice of mathematical reasoning involves a counter - in other words an encoding of time. In general, how should this encoding be implemented in the brain to account for the mental algorithms involving a counter ? A bold extension of this question may be: how do instanciations of counters in the brain relate to the general intuition of time ?

The *mind in the process of practicing mathematics* corresponds to one of the main modes of the experiencing subject that I would like to consider. The second mode is the one of the *mind in the process of meditating*. As a matter of fact, the study of meditation's effects on the brain has attracted the attention of neuroscientists amongst which S.Lazar (see for instance the article [L12] and the presentation [L12p]), who proved that meditation leads to increases of gray matter density - in other words it has a material effect on the brain. I like to account for this fact in the following way. Human minds have a tendency to '*forget*' certain beings in favor of others, meaning that they neglect them systematically in the experience - in other words these beings are in general absent from the conceptualisation of experience. The reason for this forgetfulness are the tendency of focusing on particular beings as well as the limited attention capacity. Materially this may correspond to the idea, widely present in the reasonings I propose, that neural activity may often be the trace of a mental operation used as input for other operations; in general this signifies that '*I*' am attracted by parts of my mind which correspond to already active part of the brain - which makes difficult to '*rediscover*' areas of my mind which have been inactive for a long time. Basics of meditation consist in the focalization over a phenomenon independant of any particular experience - such as the sensation of self or the mechanical act of breathing - or at least in-significant - such as the sensation of the ground in the feet. With this exercise, the activity state of the various other brain areas is rendered undifferenciated, uniform - in other words the mind is "*open*" or "*free*" because it is not naturally led to particular ideas or mental reflexes coming from an excessive focus. From this it is not difficult to see how to derive the '*scientifically validated*' positive effects of meditation, such as the reduction of tension, depression, anxiety, insomnia, increase in attention capacity, etc (which are also translated into growth and reduction of some brain areas [L12]). Since the activity of brain areas is rendered undifferenciated, beings which were forgotten may reappear: thus the association of mediation with the idea of discernment of the inside world *flickering richness* - this is, I believe, the mental stage which is represented in Alchemy by the colorful peacock.

The third mode of the experiencing subject I am interested in is the one of the *mind when acquiring reflexes, habits.* When for instance when learning how to manipulate a certain machine, I have to instanciate each time the mental algorithm corresponding to '*how to use this machine*'. With time I do not have to make this effort - in other words I do not need to think about it when using the machine - because I *know* how to use it. For me it is natural to think that while learning each instanciation of the algorithm are done in a similar way as the one I described for computing machines, when learning is over the algorithm is actually hard-wired in the brain - this could explain why its execution is then done with more ease. The question is then: provided this hypothesis, how can we account for the constitution and modulation of this (neural) circuitry ? - in particular without complete central control. I think that neural activity itself may be a marker which mechanically attracts newly created neurons. Neurons assembled this way would imitate an activity pattern and thus reproduce the corresponding algorithm in a hard-wired way. I am guessing that without a sustained activity these neurons are misled and thus stay inactive long enough to be dismantled, by the same mechanism which may modulate neural density when inactive (probably under other conditions). For similar reason, connections created this way should be close to each other - in terms of metrics, which may account for the possibility to make connections between ideas only on the short range, in other words ideas which are thematically close. The idea of a relation between (neural) activity and creation of new neuronal connections may also explain the flexibility of brain organization. For instance S.Dehaene [S17] mentions measurements done on the brain of a blind mathematician who seems to recycle neurons of the area 'dedicated' to vision for mathematics. I believe however that it is easier to think that the neurons which occupy the same area in the blind mathematician and the non-blind one derive from growth of neuronal structures out of different activity patterns - defect of visual activity and mathematics as a replacement, respectively attention to visual stimuli and then later mathematics - rather than 'specialized' neurons having to change their career. In a similar way, this relation may explain how meditation modulates neural density.

Furthermore, I hold the idea that mathematical objects are present in the unconscious mind, under one form or another, and that the practice of mathematics consists in reproducing them in the conscious mind. Assuming this is the case, what mechanisms may explain how this reproduction is possible ?

III.4.2 – Perturbations of the experiencing subject's dynamics

Some perturbations may change important characteristics of the experiencing subject, which is of interest in the study of its possible forms. I would like first to speak here about some experiences I had under marijuana some time ago, which I think may be interesting to analyse: **(i)** I had an experience similar to out-of-body experience - meaning one during which the subject perceives the world as if in a location outside of the body - in the sense that I felt as if my experience was the 'movie' of someone else's experience, one second after this person actually lived it. I think that it was still my experience, but somehow I was not able to attribute it to me. Does there exist a mechanism which integrates the experience as a whole and 'I' together ? In this case how to does the perturbation (marijuana) affects this mechanism ? **(ii)** Another time, I had my various senses disintegrated: I could perceive objects in the world through each of these senses - each of them individually was not affected - but I was not able to integrate immediately objects of different senses that I usually consider to form an object together. For instance I could touch a table, hear the sound that it makes when I hit it, or look at it, but the '*tactile table*', the '*auditory table*', and the '*visual table*' were not the same. There might be a disturbance of a mechanism creating connections between aspects of the same object: does this disturbance happen in the global neuronal workspace or in hard-wired connection ?

I think that it is possible that I both points (i) and (ii) what roots these two types of experiences is a perturbation of the sense of time and a general slowdown of activity. In fact I had another experience **(iii)** in which I felt trapped into a temporal loop, of which I had the sensation that it would never end.

Addiction consists in another form of disintegration of the self, because it bears with it situations in which some decisions that are made by 'I' without 'I' willing to make them. How to account for this disintegration ? It may be possible that multiple personalities disorders are an extreme manifestation of only this multiplicity of 'I' which is inherent to every experiencing subject. After all, '*Je est un autre*' - said **A.Rimbaud**. In fact, how to we attribute decisions and actions to ourselves ? I think that what is mysterious in multiple personalities disorders is the nature of '*personality*' and what such a term may refer to in the brain. A look into the previous question might teach us something about this.

Some perturbations on thought mechanisms may be caused partially from inside: for instance emotions affect thoughts. How exactly do they ? One property to take into account is that wether they are positive or negative, even if they are cause by specific thoughts, emotions have a global effect on thinking which does not depend on the content of thoughts themselves. I think that this property may provide a possible hypothesis on how this effect is realized: via a diffusive transmission of information rather than wired, thus outside of neural structures. Astrocytes may support this transmission of information, and provide a way through which it affects neural activity. On the other hand how to explain the conceptualization of emotions ? How can some thoughts trigger systematically a specific emotion ?

I have observed that some emotions like wrath affect our judgements about causation: when I am angry and searching for the cause of my pain or actually any mental event, I usually tend to shorten causal chains or to consider only causal relations between '*macroscopic*' objects rather than '*microscopic*' ones. This suggests overall a reduction of the capacity to make causal judgements. It does not seem to be the case for all of them: how to characterize the emotions which have this effect ? Furthermore how are causal relations encoded in the brain and how to account in it for the relation between causal judgements and wrath ? In fact think that wrath acts on attention before causal judgements, and the relation between the scope of attention and causation is clear. These intuitions may serve as constraints on the way causal relations are encoded in the brain.

In fact I have also noticed that emotions can change our conceptualisation of an experience: for instance when I was in a half-conscious state while waking up, I tend to see objects in the dark as human beings, whose positions are coherent with the form of the objects I perceive. I believe that these half-conscious states reveal a fundamental conceptualization of experiences over-ruled in conscious states which tend to priviledge animate objects over inanimate ones, for the simple reason that animate objects are more likely to be dangerous.

We may also see dreams as perturbed dynamics of the experiencing subject. I think that certain properties of dreams may relate to the integration information realized by the global neuronal workspace structure. In particular it seems that the presence of certain objects or aspects in the dreams is correlated with the occurence of events during the last day and the causal impact of these events and related objects and aspects. This suggests that this presence should coincide with a residual (neuronal) activity. Furthermore this activity may be what triggers connections between residually active areas in a chaotic way. The difference between dreams and reality would then be the action of a control on which connections are made. I think some shamanic practices as well as some transcendental operations in phenomenology - such as the ἐποχή - consist in countering an over-control of how these connections are made, and thus bringing a part of dream to the reality in a sense, opening the mind to objects which are real and present but unperceived because of this over-control. How do dreams and psychedelic states differ in these terms ?

III.4.3 – Free will

One important principled constraint on the form of the experiencing subject is free will. Of course it seems possible to conceive a subject of experience without this free will, however it is difficult to imagine such a subject entering into a relation with his or her experience without it. G.Tononi proposed an account of free will based on his *Integrated information theory* in his presentation [T21]. The baseline is the following:

1. What *exists* consists in "*maxima of intrinsic irreducible cause-effect power, at the optimal grain*". Intrinsic, because considering causation from the subject's point of view: what does exist for the subject ? The other properties specify causal properties which are meant to describe what exists for the subject. By contrast, experience from an extrinsic point of view is an experience in a world which consists in the collection of possible experiences of subjects in the same *situation*. We may associate this point of view to an abstract subject in whose point of view causal relations are absolute. Since this subject is not real, it does not make sense to describe what exists for it. By convention, anything which has causal power exists.

2. In this sense, I exist intrisically because of the causal power of 'I'. In other words I have free will from my point of view because from this point of view I can be the cause of a decision for instance between alternatives, but I do not have free will from the extrinsic point of view because there 'I' (or an extrinsic equivalent which is caused and causes 'I') is not the cause of this decision. Furthermore it is possible to exhibit this kind of change in causal structure properties on formal dynamical systems.

I shall agree with the second point, except on the following point: can the object 'I' from the intrinsic point of view be really identified with the 'I' from the extrinsic point of view ? Furthermore if we identify them in Putnam's sense, how can we exhibit this identification ? From what point of view the causes should be evaluated for this ? In fact, without definition what 'I' is, this second point shall be reduced to making sense of how free will may be compatible with extrinsic determinism, and would not apply any constraint on introspective investigation. About the first point, I agree with the intrisic approach, but not necessarily with the characterization of what exists intrinsically in general. This kind of position presents in fact severe epistemilogical difficulties which I discussed in my other paper [G20].

Here I would like to offer another point of view on this matter, with a special care for the difficulties of the second point, which I hope can shed light on how the constraint of free will may steer introspective investigation.

It seems clear to me that from any point of view, 'I' is related to causation, for some of my actions are not derived from my will, for instance because they fast reactions in situations when I do not have time to ponder on what action I should take. They way I differenciate, from my own point of view, these actions from the actions which derive from a decision, is causal: 'I' is not the cause of this action, but is caused by a part of my '*extended self*' (probably a similar notion to Damasio's protoself), which may be defined as '*what each of my experiences bears with it*'. This extended self is thus identical to 'I' in Putnam's sense, an identification which is reflected in the language: we usually designate both 'I' and my extended self by '*me*'. Then if what allows this distinction is expressible in causal terms, it is possible that the definition 'I' should also be so.

I propose the following definition: 'I' is a concept constructed as the common cause of events which are are actually caused by an event in my extended self, for which I do identify a pragmatically ultimate cause in what I designated above as the *place of 'I'*. I think a similar definition may hold from another subjective point of view; however would it hold from the extrinsic point of view ? It seems clear pragmatically that they can be identified across subjective points of view, but it is not how this identification is done.

This definition calls for other questions, which does not necessarily presupposes the definition: by what mechanism do I attribute the cause of a certain event to myself ? Furthermore how to formalize *precisely* the notion of cause involved in the definition ? The first question can be investigated through the wider question of: *how do I attribute causes to an event*? as well as through particular modes of the experiencing subject (including the simulation of mental computing machines). The second one by progressive refinements of the naive notion of cause in counterexamples (differenciating for instance the causal role of respectively reasons for a choice and 'I'). We can also examine particular classes of mental events: for instance doe integration event involved in the local instanciation of a mental algorithm come systematically with an attribution of cause to 'I' ?

Beyond natural implications of the question of free will for law, there are implications of understanding what 'freedom' is on the general form of the experiencing subject, because it is reasonable to think that one of the fundamental desires of human beings is freedom (the realization of which may vary for different '*cultures*': for instance some may search for freedom in their mental representations, others their actions in the physical world), certainly affecting the way they think. This will for freedom may act as a constraint on introspective investigation.

I believe that *being free* is not about the possibility of an arbitrary choice (which would be absolutely undetermined), or the simple possibility of a choice, for in this sense there is no situation in which we are not free: *Je peux toujours choisir, mais je dois savoir que si je ne choisis pas, je choisis encore.* - said **J.-P. Sartre**. It is rather about the possibility not to predict, the possibility to hope - which is a necessity of life: without any form of hope, what reason would there be to live ? More formally, the largest the space of seemingly reachable possibilities, the more free we fill. And in fact we feel the most free when truely travelling, because then we are immersed in the unknown. Locally in time, we tend to be attracted to cognitive *places* where contingency is the highest. This means not only the greatest number of possible choices, but also the most ease to switch from one to another if a certain event makes obsolete the conditions of a first choice.

The question is the following: how to formalize these intuitions and account for them in the model of the experiencing subject ? Some difficulties appear here: for instance, what should we count as alternative possibilities ? **(i)** Do we only count actions possible in the present moment (how to define the present moment ?!) or do we count foreseeable possible effects in order then to choice the appropriate action for this effect ? The reality of the relation between an action and an effect does not matter here, and this adds up on the difficulty. Furthermore, effects may be over variable time intervals. **(ii)** The evaluation of ease to modify a choice aftermath is dependant upon the situation, and relates to high level notions such as courage.

Digression: *The considerations on causation here call questions of the following kind: is it possible for an event to have no cause ? Is it possible for an event to have an actual cause across time ?*

III.4.4 – On information integration

Integrated information theory formalizes information integration in a particular way, however there are other possibilities to do so. Furthermore, this formalization is concerned mainly with statical experiences. Here I would like to mention other ways to think about information integration.

Integration in dynamical experiences: In particular, statical experiences have the advantage that their conceptualization is relatively stable (when well defined): in principle we can talk about objects which appear in this conceptualization, and attempt to explain why some patterns are object and why others are not. However we may also consider more dynamical experiences and situations in which an object which was not present to the mind *appears* to it: we need to develop methods to introspect on these. For that there are various types of integrations which we should distinguish objects which are actively created and objects which are passively perceived as such (in particular in situations in which the conceptualization of an experience or set of experiences is simplified, structured, organized).

The second type of apparition may be related to causation: as causal relations '*direct*' the focus of my attention from one term of the relation to the other, a pattern which is causally stable [G20] corresponds to a pattern of attention which groups together the elements of the pattern, as they are all active in the 'same time', which makes possible their grouping by neuronal connections.

The first type of apparition was mentioned already above, when multiple instanciation of a mental algorithm result in a hard-wired version of this algorithm. In this direction, we may think about the 'computing machine' supported by the global neuronal workspace as an '*integrating machine*' meant to control necessary connections between parts of the brain, and the consequent creation of objects. We can notice here that the idea of a balance between *functional segregation* and *integration*, rooting the formalism proposed by G.Tononi [EST94], is in a sense realized by this integrating machine, suggesting another way to formalize it. Furthermore this of thinking about it seems coherent with phenomena steering the relation between integration and consciousness, such as the division of consciousness in split-brain patients, with a change in the interpretation of the conceptualization split-brain patients's behavior as the division into two conscious minds. In fact after brain split the integrating machine remains probably unique: only its tape is divided into two parts, between which it may switch. However because of the brain split, brain mechanisms involved in integration - provided their nature - are limited to one of the two *domains* of the 'tape' where the machine 'head' is present, and no integration can happen between these two domains. On the other hand the integrative capacity of the machine is left unchanged in each of the domains. From an exterior point of view, since integration probably appears in a similar way, the best conceptualization of the split-brain patient's behavior induces the idea of two distinct minds (similarly to multiple personality disorders). On the other hand this is dependant upon what we would like to call '*one mind*'; while the 'tape' of the machine is divided, the 'head' is probably not: which one defines what *one mind* is ?

I would like also to mention here that one way to combine and co-form the approach of integration in causal terms and in terms of computing machines would be to evaluate the quantity of information integration that a system structured like the dynamical we can define out of introspection on the structure of the experiencing subject and prove that it is higher than other simpler systems. Also, another way to look at integration is to consider situations in which the mind has to ensure information separation and *dis*-integration: for instance when separating two concepts which have been erroneously identified or dismantling a certain habit (in this case we should account for the difficulty of getting rid of a habit). I think that some disintegration can be done via inhibition (in fact O.Houdé [H19] for instance has shown that inhibition is part of human intelligence as much as the execution of complex algorithms): preventing systematically the use of some neuronal connections expose them not to be maintained, and thus ultimately destructed. However as for integration there might be multiple ways to approach it.

What about astrocytes? I find surprising that these cells are not taken into account in the theorisation of consciousness, while they have been proved to have an important role in information processing in the brain. As a matter of fact, they have also been related to consciousness in experimental research (see for instance [R02]): in particular they are involved in information integration and perturbations of consciousness are correlated with perturbation of astrocytes' dynamics. In [PF09]:

"*The neuron is, computationally, a filter that converts analog to digital-like information, while astrocytes can be described as being like a hub able to integrate patterns from around 100,000 to 140,000 synapses. They integrate excitatory inputs received from neurons connected to their tips.*"

As a matter of fact, neurons can not be thought to support every information processed: otherwise, how to differenciate which neurons support conscious and unconscious processing ? As a consequence it is reasonable to think that other cells matter regarding conscious experience. Furthermore, since astrocytes can be considered as a contact point between neurons and the remainder of the brain - through which elements used to repair neurons are received - it makes sense to think that intruders such as marijuana should affect consciousness through disturbance of the dynamics of astrocytes. I think it is reasonable to put forward the hypothesis that while neurons only contain information under the form of signals and neuronal structures implement algorithms, astrocytes realize integration between these bits of information and with mental states, relatively to their state.

It is possible that the region of the brain called *global neuronal workspace* has a high concentration of astrocytes which may be involved in the attribution of 'functions' to neurons close to the 'machine head' I talked about above, in order to execute simple operations, inhibit connections between regions supporting local 'routes' while instanciating some algorithms, or attribute this route function. The machine would act directly on astrocytes and indirectly on neurons.

III.4.5 – Some other thoughts

I mentioned above the possibility to introspect on mechanisms of introspection itself. There are at least two things to consider: the action of 'highlighting' a certain part of the conceptualisation of experience (for instance a signle object), making it more 'present', and they translation of what I perceive in me into words (wether these words consist in mathematical objects or words of the natural language). It seems that the way it is done in the mind consists in 'projecting' towards the area I am in some articulations of words, in a partially random way; this projection is followed each time by the action of a mechanism which compares this articulation to the unconscious content of this area, until this mechanism judges the articulation close enough to this content, at which point this closeness is signified and the search comes to an end. How is this process realized in the brain ? More generally we could consider properties of the language and its creation as constraints for introspective investigation.

I also believe that it would be particularly interesting to think about how the global neuronal workspace structure may be constituted during the brain's development and what constraints this add on its actual structure. In other words: how does consciousness appears and grow ? My hypothesis here (that I consider to be even more speculative than the remainder of this paper) is that this structure is constituted in two stages: **1.** centralization of information to be integrated to a single hub with a retroaction loop checking error in the function integrating information. **2.** it is out of the development of this function that the global neuronal workspace is created, as the amount of information to integrate becomes higher and integration delegated and hierarchized.

References

[B66] **R.Berger.** *The Undecidability of the Domino Problem.* Memoirs of the American Mathematical Society, 66 (1966).

[BP13] **M.Bitbol and C.Petitmengin.** *On the Possibility and Reality of Introspection.* Mind & Matter Vol. 14(1), pp. 51–75

[C18] **D.Chalmers.** *The meta-problem of consciousness.* Journal of Consciousness Studies 25 (9-10):6-61 (2018).

[C20] **D.Chalmers.** *Debunking Arguments for Illusionism about Consciousness.* Journal of Consciousness Studies 27 (5-6):258-281 (2020).

[D13] **S.Dehaene.** *The brain mechanisms of conscious access and introspection.* Neurosciences and the Human Person: New Perspectives on Human Activities. Pontifical Academy of Sciences, Scripta Varia 121, Vatican City 2013.

[S17] **S.Dehane.** *A close look at the mathematician's brain?* video link.

[D15] **D.Dennett.** *Why and How Does Consciousness Seem the Way it Seems?* In T. Metzinger & J. M. Windt (Eds). Open MIND: 10(T). Frankfurt am Main: MIND Group.

[D16] **D.Dennett.** *Illusionism as the obvious default theory of consciousness.* Journal of Consciousness Studies 23 (11-12):65-72 (2016)

[D78] **D.Dennett.** *Where am I ?* Brainstorms. MIT Press (1978).

[F16] **K.Frankish.** *Illusionism as a theory of consciousness.* Journal of Consciousness Studies 23 (11-12):11-39 (2016).

[G20] **S.Gangloff.** *A formal window on phenomenal objectness.* Unpublished.

[Giles] **H.A.Giles.** *Chuang Tzu: Taoist philosopher and Chinese mystic.* Allen and Unwin, 1926, London, p. 47.

[HM10] **M.Hochman and T.Meyerovitch.** *A Characterization of the Entropies of Multidimensional Shifts of Finite Type.* Annals of mathematics, Vol. 171 (2010), Iss. 3, pp 2011-2038.

[JR92] **B.Josephson and B.Rubik.** *The Challenge of Consciousness Studies.* Frontier Perspectives 3 (1):15-19 (1992).

[KT20] **J. Kleiner, S. Tull.** *The Mathematical Structure of Integrated Information Theory.* Preprint. pdf link

[N94] **T.Nagel.** *What is it like to be a bat ?* The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.

[IIT] **M. Oizumi, L. Albantakis, G. Tononi.** *From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0.* PLoS Comput Biol. 2014 May; 10(5).

[P20] **R.Penrose.** *Mathematics, Mind and Consciousness.* video link.

[R71] **R.Robinson.** *Undecidability and Nonperiodicity for Tilings of the Plane.* Inventiones Mathematicae, 12(3), pp. 177–209 (1971)

[L21] **C.List.** *Many-worlds theory of consciousness.* Preprint. pdf link

[G15] **T.W.Webb, M.S.A. Graziano.** *The attention schema theory: a mechanistic account of subjective awareness.* Front Psychol. 2015; 6: 500.

[HT19] **A.Haun, G.Tononi.** *Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience.* Entropy 2019, 21(12), 1160.

[Turing] **A.M.Turing.** *Computing machinery and intelligence.* Mind, New Series, Vol. 59, No. 236 (Oct., 1950), pp. 433-460.

[K20] **J.Kim.** *How can my mind move my limbs?* Philosophic Exchange, Vol. 30 [2000], No. 1, Art. 3.

[Meillassoux] **Q.Meillassoux.** *After finitude: an essay on the necessity of contingency.* Bloomsbury Publishing, 2009 - 160.

[Searle] **J.Searle.** *Minds, brains and programs.* The behavioral and brain sciences (1980) 3,417-457

[P67] **H.Putnam.** *The nature of mental states.* in Readings in Philosophy of Psychology, Volume I, Harvard University Press (2013).

[S20] **L.Shapiro.** *Multiple realizations.* The Journal of Philosophy 97(12): 635-654.

[P60] **H.Putnam.** *Minds and machines.* In Sidney Hook (ed.), Dimensions of Minds. New York, USA: New York University Press. pp. 138-164 (1960)

[H45] **J.Hadamard.** *The psychology of invention in the mathematical field.* Princeton University Press, 1945.

[P13] **H.Poincaré.** *The foundations of science: Science and hypothesis, The value of science, Science and method.* The Science Press, 1913.

[S60] **G.Sperling.** *The information available in brief visual presentations.* Psychological Monographs. 74 (11): 1–29.

[SA16] **M.Amalric, S.Dehaene.** *Origins of the brain networks for advanced mathematics in expert mathematicians.* PNAS, May 2016, 113(18), pp. 4909-4917.

[L12] **B.K. Hölzel, J. Carmody, M. Vangel, C. Congleton, S.M. Yerramsetti, T.Gard, S.W. Lazar.** *Mindfulness practice leads to increases in regional brain gray matter density.* Psychiatry Res. 2011 Jan 30; 191(1): 36–43.

[L12p] **S.W. Lazar.** *How meditation can reshape our brains.* video link.

[T21] **G.Tononi.** *Integrated information theory and its implications for free will.* video link.

[EST94] **G.Tononi, O.Sporns, G.M.Edelman.** *A measure for brain complexity:relating functional segregation and integration in the nervous system.* Proc. Natl. Acad. Sci. USA, 1994 May 24; 91(11):5033-7.

[H19] **O.Houdé.** *L'Intelligence humaine n'est pas un algorithme.* Odile Jacob Psychologie, April 2019.

[R02] **J.M.Robertson.** *The Astrocentric Hypothesis: proposed role of astrocytes in consciousness and memory formation.* Journal of Physiology-Paris, 2002, 96(3-4):251-5.

[PF09] **A.Pereira, F.A.Furlan.** *On the role of synchrony for neuron–astrocyte interactions and perceptual conscious processing.* J Biol Phys. 2009 Oct; 35(4): 465–480.