Evolutionary theory of consciousness: a comprehensive model

Sergio Graziosi

Social Science Research Unit, UCL Institute of Education, University of London, London, United Kingdom

Correspondence:
Sergio Graziosi,
Social Science Research Unit,
UCL Institute of Education,
University of London,
18 Woburn Square,
London, WC1H 0NR, UK.
s.graziosi@ioe.ac.uk & s.graziosi@ucl.ac.uk

Keywords: consciousness, evolution, learning, phenomenal experience, hard problem.

ABSTRACT

This paper uses a theoretical and deductive approach to address the hard problems that arise when trying to explain what consciousness is and how it works. The starting question is: why did consciousness evolve? I argue that it is possible to propose an initial answer to this question by hypothesizing small and incremental additions to primordial, genetically-encoded and fixed reactions to perceived stimuli. This approach follows the well-accepted, prototypical pattern of evolution as driven by natural selection. Organisms that rely on learning new behavioral strategies need to constantly evaluate sensory information in order to gradually understand the world they live in: they actively evaluate perceptions, so as to build a map of relations between present/past perceptions and link them to internal motivational states; this ability is essential to predict the consequences of possible actions and to inform future decisions. Building on this principle, I propose a conceptual model of consciousness, subdividing it into distinct computational modules and functions that may be identified and studied separately. This same subdivision allows for significant gains in conceptual clarity: it makes it possible to address the differences between most mainstream theories of consciousness by showing how many theories appear to model only subsets of functions; for example, some theories focus on perception, while others model awareness and attention, thus they inadvertently fail to describe the full picture. At the same time, the full model proposed here allows us to explain the existence of the problem of phenomenal experience: I will show why any system that follows the proposed signal-processing principles would necessarily find the same difficulties that we encounter when we introspectively analyze the nature of perception.

Introduction

Can consciousness be defined, understood and modeled with a reductionist approach? In this paper I will show that it can, with the essential aid of the evolutionary lens, if one carefully avoids modeling only a subset of its functions. Thus, the starting point of this ambitious endeavor will follow the path indicated by Searle (2013): in order to understand what consciousness is, what it does, and how it does it, the most promising question one needs to address is *why did consciousness evolve*? In the process of answering this question we will gain valuable insights on what consciousness does, allowing us to evaluate some current empirical

evidence and the theories that guide its collection and interpretation. The result is a conceptual theory that allows us to build a complete model of consciousness; I will call it Evolutionary Theory of Consciousness (ETC). I should point out that ETC is neither the first nor the most developed theory of consciousness that starts from an evolutionary point of view. However, ETC addresses a wider range of conceptual problems, and does so by taking an explicitly algorithmic stance. Thus, it is specifically designed to aid the progress of Cognitive Neuroscience. For an overview, and a still not comprehensive summary of the approaches available, including but not limited to the evolutionary perspective, see Weisberg 2014, Dehaene 2014, Revonsuo 2009, Blackmore 2006, Bachmann 2000, Humphrey 1998 and Baars 1997. The reader should also refer to Metzinger (2003).

Before developing ETC, however, it is worth discussing the general aim, the intended role of ETC and the limitations that directly stem from the chosen approach. Brain sciences in general, and cognitive neuroscience in particular, all share a common goal: they try to build a comprehensive explanation of how our brains work. The key word is *comprehensive*: to be convincing, such an explanation needs to account for the detailed mechanisms at molecular, cellular and intracellular levels. However, the task cannot be completed without also considering the systems level, the overall description of how different brain regions interact with one another and finally building an explanation of the high-level psychological phenomena, as observed from a third-party perspective, but perhaps more importantly, as experienced directly by each one of us. In short, the overarching aim is to build an explanatory theory that could convincingly show how the brain mechanisms (subject to empirical enquiry) may eventually produce what we call a mind. The difficulty of this task cannot be overestimated, but the progress made so far does allow us to identify and isolate a range of different problems that would need to be solved. In the following paragraphs, I will briefly discuss the main ones. This short and not exhaustive list will serve two aims. First, it will allow me to explain why cognitive neuroscience would benefit from a careful and rigorous analysis of a handful of classic philosophical problems. Second, I will be able to show how ETC tries to aid this operation, with the final aim of providing a compass to navigate the space of more detailed explanations.

Assuming that we will someday gather a complete picture of the molecular, cellular and intracellular brain-mechanisms, one big challenge will certainly entail building the correct bridges between different scales. Although this effort will be strictly empirical and reductionist, its description already exposes a treacherous problem: the apparent tension between reduction and emergence (see for example Damper 2000 and Sawyer 2002). Most of us hope that producing the mechanistic explanations will help resolving this tension, but we need to remain aware of the problem. Furthermore, the final explanatory bridge – the one between a comprehensive mechanistic explanation of the brain and direct experience of our own mental life – is widely thought to require some significant and extraordinary additional efforts. The detail of this conundrum is explained in part one, but for now it is useful to point out that bridging this gap is indeed part of the challenge that brain sciences are trying to address. Thus, it will be useful to produce an initial idea of what such a bridge may look like. ETC aims at doing exactly this: instead of brushing aside the known philosophical problems that inevitably cluster around cognitive neuroscience, it tries to use them in a positive, propositional way. By concentrating on the obstacles, ETC tries to reduce the space of possible solutions, thus providing a high-level guide for further work.

At a more practical level, it is widely recognised that empirical efforts in the study of the brain face the challenge of finding useful ways to catalogue, classify and synthesize an enormous amount of data, in order

to produce intelligible models that allow interpreting and predicting the behaviour of the most complex system known to humankind: the human brain. Understanding how to identify the correct abstractions that could be employed by such models is not at all trivial, and is probably the main unifying topic in the recent volume *The Future of the Brain: Essays by the World's Leading Neuroscientists* (Marcus and Freeman, 2015). The central message of this stimulating book is that in order to reap the benefits of the latest, large-scale efforts in neuroscience, it is necessary to produce suitable interpretative frameworks. Bottom-up, unguided analysis of big-data will probably produce valuable insights on its own accord, but we also need to develop new, well-substantiated guiding principles that may speed up the identification of the more useful levels of abstraction. In this context, it is worth noting that producing top-down abstractions is an extremely risky exercise, especially if the aim is to identify classifications that may eventually match and correspond to the outcomes of the bottom-up efforts. To give just an example of this difficulty, the study of attention is paradigmatic. In everyday life, we all more or less agree about what attention entails, however, when it comes to isolating and describing the fine-grained mechanisms that produce attention, difficulties emerge already at the level of definitions: no one can, a priori, produce a definition of attention that would be consensually accepted by all. Even within the considerably narrowed-down field of visual attention, as studied in experimental psychology, finding a way to unify disparate and seemingly antagonistic explanations is not trivial at all; see for example Vecera et al. (2014). Thus, the difficulties are symmetric: on a bottom-up approach, one could use a priori methods to identify meaningful regularities. However, it is easy to expect that some of these regularities will actually turn out to be spurious (a side effect of the vast data analysed: the larger the data-sets, the more likely it is that some patterns will emerge by pure chance). Some others will certainly be authentic, but on their own accord may tell us little or nothing about how to link them to larger scale explanations. Conversely, top-down approaches easily make plenty of conceptual sense, but might not be accurate descriptions of how the real mechanisms work.

To summarise: having a reliable theory that could guide the interpretation of bottom-up data would be very useful. Unfortunately, reliability in this field is impossible to establish a priori. The difficulty is heightened by the fact that what we are trying to explain (how brains produce minds) is inextricably linked to our own direct perception of having a mind. Thus, the problem of subjectivity, and therefore consciousness, is central to all the disciplines that study the brain, and to cognitive neuroscience in particular: an acceptable unifying framework will necessarily need to answer the question of how neural activity generates subjectivity.

For all these reasons, ETC tries to provide a high-level, overarching theoretical compass that may aid empirical research. By focussing on the seemingly intractable problem of consciousness, it tries to turn the problem into a resource: finding a convincing theoretical explanation of a difficult issue (one that is part of the relevant explananda) should help generating more detailed theories, and thus shorten the distance between theory and practice. On the other hand, the method through which this theory was produced was carefully shaped by the sharp awareness of the limitation associated with such an endeavour: purely theoretical hypotheses cannot be considered scientific *a-priori*.

To minimise the risk of creating yet another theory that tries to be general, but is only reliable on limited and selected domains, the following countermeasures were adopted. First: such a top-down theory, will have to be built in a deductive way. Thus, it is important to select the correct deductive framework. In this case, the choice fell on our theoretical understanding of evolution. This is justified by several observations: our brains have been primarily shaped by evolution, therefore the known facts on how evolution proceeds should provide the appropriate conceptual guide to start the journey. Furthermore, evolution is the most solid

theoretical knowledge associated with biology, making it the most appropriate set of applicable deductive rules.

Second: the result should be able to accommodate the widest possible amount of empirical evidence (see part two). Thus, it needs to be highly abstract and avoid falling into direct contradiction with the empirical data that supports the already existing theories.

Third: however, in order to be useful, such a theory would need to avoid being too abstract. A theory that can accommodate any and all evidence is not a scientific theory. Therefore, the final result will need to make some testable prediction. However, since the aim is to provide an overarching framework without over-estimating the reliability of theoretical inference, it is important not to get carried away and instead, limit predictions to the essential. On the one hand, this ensures that the effort is not vain or pseudo-scientific, on the other, it does not restrict the applicability of the theory produced.

The discussion above thus provides a guide for the reader, and explains why different approaches are utilised in the separate sections. **Part one** is deductive: it uses an established theory to produce a first, very abstract hypothesis. Since the aim is to produce a reductive model, the apparent limits of deduction, as identified by Chalmers (1995) are briefly discussed. **Part two** utilises two different approaches. It initially focuses on the known conceptual problems: the aim is to verify if the current hypothesis may indeed produce reasonable answers – across the whole paper, philosophy of mind is used to identify the known problems and to verify if ETC does indeed seem to be *useful*. Having answered positively, the symmetric exercise is attempted: I will verify if ETC can work as an interpretative framework of bottom-up, empirical, pre-existing and alternative scientific theories. **The conclusion** returns to the big issues: in the empirical domain, it shows how ETC can indeed produce testable predictions. On the theoretical side, we will see that ETC may help answering the first of all bridging questions: the problem of consciousness itself.

At the present time brain sciences still look like a collection of isolated disciplines, and the corresponding body of knowledge is comprised by a multitude of theories and proto-theories, fragmented and isolated within their own boundaries. The short list of “unsolved problems” provided by Adolphs (2015) is grim, but importantly, it does make two positive contributions. First, Adolphs reminds us that “understanding the brain” requires building explanations for what we experience as our mind. Whether such “folk psychology” is ill defined, somewhat illusory or flat-out wrong is irrelevant: an explanation will not foster consensus unless it is able to account for our naturally occurring intuitions. Second, he advocates focusing on algorithms and taking a comparative approach (across species and across abstraction levels). In this context, ETC embraces the suggested method, builds an algorithmic picture of what consciousness does, and provides the necessary foundation to start comparing the functions associated with consciousness across different species. Thus, although (or because it is) highly theoretical, ETC tries to provide one of the missing ingredients in current cognitive neurosciences: a theoretical framework that can help bridging sub-disciplines and levels of enquiry.

Part one: building an evolutionary theory of consciousness

In very practical terms, one could say that consciousness is the ability to know who we are, what we are doing and, at least to some extent, why we are doing it. This definition works quite well if we use it to guide our use of language in ordinary life, but it does not help much when one tries to dissect the philosophical implications of such ability. It is however a useful starting point, because it does define a function that is not mysterious in practical or evolutionary terms: it should be quite uncontroversial that knowing what you are

doing and why you're doing it can facilitate survival, especially when this ability is used to supervise, fine-tune or complement a pre-existing set of automatic and/or instinctual responses.

Thus, I will use this provisional definition as the guiding light of my reductionist attempt. The aim being a detailed model of how organisms implement this function: the starting point will be a bird's eye view account of the basic and well understood mechanisms that support perception. This will provide the platform on which to apply basic evolutionary considerations: the process in this phase will be one of strict deduction, applying accepted general rules on how evolution operates (extending, redeploying and modifying existing structures/mechanisms), with the aim of identifying possible functions that may evolve from primordial modes of perception. This phase accounts for the "Evolutionary" side of ETC and is useful to identify possible/plausible mechanism by eliminating all the theoretically viable ones that would not, or could not possibly evolve from primitive perception systems. This phase will thus stop at a high theoretical level, just after identifying a function that is bound to enhance fitness and may be useful to start explaining the many riddles that surround consciousness.

Scientific method is frequently strictly reductionist: it subdivides the object of inquiry into smaller components, defines how each part works, how they interact, and uses this understanding to create a model of the whole. However, on the subject of consciousness, this approach has been so far fairly fruitless, and its usefulness has been effectively (and usefully) questioned. In particular, Chalmers (1995) has highlighted the reasons why this approach may be unable to clarify both what consciousness is and how it works. In his seminal paper Chalmers was able to separate which aspects of consciousness can be explained in reductionist terms from those which cannot: he thus isolated a number of supposedly "easy" questions from a "hard problem" that may be empirically intractable. His own definition of the core problem is wonderfully synthetic and, most importantly, *useful*:

> *If someone says "I can see that you have explained how information is discriminated, integrated, and reported, but you have not explained how it is experienced", they are not making a conceptual mistake. This is a nontrivial further question.*

Defining this "nontrivial further question" is useful because, although the whole of Chalmers' paper challenges the reductionist approach, it also empowers it in two ways. First, it isolates the parts of consciousness that can be tackled with standard methodologies ("how information is discriminated, integrated, and reported"); this is useful as it can (and does) empower empirical research. Secondly, and more importantly, it also highlights the issue that does require an additional effort – the "hard problem" further defined as follows:

> *Why doesn't all this information-processing go on "in the dark", free of any inner feel? Why is it that when electromagnetic waveforms impinge on a retina and are discriminated and categorized by a visual system, this discrimination and categorization is experienced as a sensation of vivid red?*

Chalmers, by isolating the problem that seems impervious to a reductionist approach, has in fact made the first and most important step in making it tractable. This paper is a direct consequence of the clearly defined challenge set up by Chalmers: to be credible, a model of consciousness needs to address two somewhat separate domains. First, one needs to build a model of simpler and tractable functions that, taken together, explain the full consciousness phenomenon; once this is done, the reductionist work is done: such a model

should be useful to make predictions (given the particular brain damage suffered by this patient, is it possible that she is still conscious?), generate new hypotheses (could this part of the brain be responsible for this particular sub-function?) and, in general, satisfy all the strictly scientific questions. However, Chalmers' additional question, on why the computations described in the model do not happen "in the dark" (for brevity, I will refer to this question as the "hard problem", or the problem of phenomenal experience – PE), does need to be addressed separately: one needs to check if and how the full model (or part thereof) does explain the existence of PE (following the common convention, I will call this explanation the explanatory bridge). This second, philosophical question will be addressed in the third part of this paper, once the full ETC model has been described.

To rise to the challenge, it is useful to start by enunciating what we already know: most animals come with sensory systems that collect information about the environment, encode it into nervous signals, and send it to the brain for further processing. This is essential for survival, as it allows organisms to react appropriately to ever-changing circumstances. We also know that crude sensory information is processed along the way, so that significant features are extracted from the raw signal in a progressive manner. It is also frequently accepted that the separate sensory pathways contribute to building and maintaining a quasi-real-time, ever-changing internal representation of the surrounding environment, along with the sensing organism itself. How this model of the world and self is built and described is, for the time being, irrelevant; the reason is simple: the general architecture of ETC is and has to be agnostic on the detailed computations; my current aim is to draw on principles that can be considered accepted across the fields, and see how they allow sketching the evolutionary history of perception first, and consciousness thereafter. In order to follow the rest of this section better, the reader should try to provisionally espouse this "agnostic" stance: if specific references or allusions to a particular framework can be found in the discussion that follows, please consider them as the accidental result of the difficult attempt of making abstract thought clear and easy to follow.

For now, it is enough to say that a model of the world and self is required, or at least useful, for the normal functioning of most animals that are able to move, allowing intentional actions in a way that is certainly useful for survival. More complex behaviors, such as hiding when a potential danger is approaching, will require more complex representations, including a computational strategy that allows to evaluate the contents of this representation, select a suitable hiding place, and trigger the actual response. Among relatively simple organisms such as a zombie-like lizard (that is: an imaginary lizard for which all signal processing goes on "in the dark", a lizard that does not have PE. For a review of the zombie concept, see Kirk 2012), a genetically predetermined reaction can certainly be effective, even if it is relatively rigid. Such a lizard may be born with prototype representations[1] of dangers and hiding places: when a danger is detected, it will match these predetermined "concepts" with incoming sensory information and react accordingly. There is little or no doubt that some animal behavior roughly follows the above description. The fundamental point is that although this system allows such animals to effectively respond to predictable needs, it is also quite limited. Should the lizard find itself in a new environment that allows for new effective ways to hide, only a genetic mutation that describes a corresponding prototype would allow exploiting the newly available

[1] As hinted in the main text above, I will intentionally avoid specifying in detail what exactly such a prototype (and/or "concept") is, how it may be represented, encoded and matched to incoming stimuli. This is because ETC does not aim to be a theory of perception, in fact it may and should be used as an additional theoretical layer useful to compare, contrast and inform other theories –including non-representational ones. In part three, for example, I will discuss how ETC may accommodate either Information Integration or the Predictive Coding theories.

strategies. A better way to maximize fitness would be to learn how to hide in new ways; but in order to do so, the following will be needed: first of all, the lizard will need to have some new, more abstract prototypes, which allow the internal model to represent suitable causal chains. In the case of hiding, this can be built by associating three fundamental prototypes: a danger, the need to hide and a hiding place. Thus, both the danger and the hiding place prototypes become increasingly effective the more conceptual they are: for example, the initial prototype of a hiding place may have been a suitably sized fissure between rocks. Making it completely conceptual transforms it to: any place/position that is outside the visual field of dangerous creatures. Applying this more generalized concept allows the animal to hide beneath the sand or around a corner, opening up a wide range of effective responses.

However, an animal that comes with these "generalized" concepts and strategies will need to learn how to match raw sensory stimuli to the appropriate representations (danger, hiding place), the consequent "need" (hide!) and the required reaction. Crucially, the more the prototypes are conceptual, the less a-priori assumptions on how to match concepts with stimuli are admissible: a danger may be a large moving animal, or could be a small competing lizard. Hence, if an organism is able to learn how to classify and react to different stimuli, it necessarily needs some specific computational abilities, and upon inspection, it turns out that the minimal set of these abilities coincides with most of what we normally describe as consciousness; it also allows looking at the hard problem from a new perspective.

Let me go back to what we know: so far we had signal reception, transduction and transmission. I'll call this **Level 1**, where some basic data processing may already start[2]: edges are enhanced already in the retina (Levick, 1967), temperature sensors react to temperature variations but also to potentially harmful constant temperatures (Saito et al., 2011), chemical receptors have adaptation systems (Kurahashi and Menini, 1997) and so on. **Level 2** is where fundamental properties of the incoming signals are extracted, creating progressively conceptualizing what is being perceived. For example, a collection of edges that move in a tightly organized way may be classified as a single object. After this, sensory information coming from separate pathways (including interoception) needs to converge and unify (**Level 3**). This is necessary, for example, to match visual information with audio signals, so to identify the source of a given sound. We already know that a zombie-like lizard, which has only predetermined reactions, needs to have a system of this kind. However, a more complex animal, not born with predefined correspondences, will necessarily need some specific systems that allow to gradually learn how to appropriately match stimuli with needs and actions. But what would these systems be? To answer this question, I will provocatively ask another one: what are fear, anger, hunger, lust and tiredness? Within the current description, they can be described as pre-defined (genetically encoded) dispositions that we are born with, they are physical states that embody our *adaptive purposes*. Via development and accumulated experience, one will learn how to match them with the appropriate stimuli (including internally generated ones), as well as what reactions are appropriate in response to the activation of different goals. To do so, we need a system to store previous experience, and a system that is able to evaluate different matching between perceived situations, experienced needs, and appropriate reactions. The ideal way to do so would be to store the full stream of sensory data and use all new information to re-evaluate it in full. However, evolution doesn't work like this, it needs to build new functions incrementally, modifying and extending existing systems. Evolution also favors efficiency, but the

[2] As before: the aim here is to sketch a broadly defined conceptual framework. I will subdivide the pathway from perception to consciousness in four conceptual levels. This subdivision is arbitrary: it is useful to make the framework understandable; the four levels are used to help the reader follow an otherwise abstract argument.

ideal system is clearly wasteful, as it requires maintaining huge amounts of redundant information that will be expensive to preserve and constantly re-evaluate. The more economical way to obtain almost the same result would be to have an Evaluation Module (EM - the first one to appear at the higher conceptual level: what I will describe as **Level 4**) that assesses what is present on **Level 3** (the unified stage) and is used to decide what to do with the new information detected. Some options are[3]:

a. This information is new, possibly significant: it should be stored and re-evaluated in the future. This may happen when we see an unknown fruit, which may or may not be edible.
b. This information is understandable, it may be useful and it is worth remembering. For example, when we notice that a product we like is for sale.
c. I know what this information means, and need to react accordingly. As when a car is approaching and might hit us.
d. This information is not new and not worth remembering. As when we enter our home and everything is as expected.

This starts as a simple classification task, and the zombie-like lizard may be able to perform b), c), and d) but arguably not a). In other words, in order to evolve this "advanced" module, little modifications are required: memory is already present in the lizard, as it surely can remember where sources or food are; what is needed is a new way of classifying information so to apply the "possibly significant, to be remembered and re-evaluated" flag, and to be able to store this information in a suitable fashion. I argue that this evolutionary step is very useful, as it allows enhanced adaptability, but it is functionally straightforward: it can be seen as a small and incremental addition to systems that are already present. However, it is tremendously significant, because it provides a direct and definitive explanation of why the hard problem exists (see conclusions).

Before concluding this section, it is useful to summarize the main theoretical considerations, the proposed consequences, and the main resulting working hypothesis that will then be further refined by looking at existing theories and scientific evidence. First, I claim that it is reasonable to assume that at some stage during evolution, some organisms have started to react to stimuli in stereotyped, genetically predetermined ways. Second, I propose that this strategy may have been enhanced by allowing an organism to react to certain situations in less predetermined ways: this would permit generating adaptive behaviors without the need of a corresponding genetic mutation. However, a mechanism of this sort can only exist if it does indeed promote the emergence of new *adaptive* behavior, or, in other words, requires the ability of learning (to discriminate and remember what works and what doesn't). This can be obtained via two extensions: the first one is the progressive abstraction of motivational states (for example "I need to hide into a small cavity" becomes "I need to be outside the visual field of predators"), the second one is an evaluation module that allows to decide what sensory stimuli are worth remembering, so that learning can happen.

As a result, we find ourselves with a list of the very basic components that may allow us to know "*what we are doing and, at least to some extent, why we are doing it*": the list includes a model of the world, a model of the self, motivational states, and a minimal system to link the two. The working hypothesis therefore is: conscious organisms are those that are able to learn new, not genetically encoded behaviors. Learning where sources of food are to be found can be a genetically predetermined ability, and is not part of the working hypothesis: I am specifically addressing the ability of reacting in new, unprecedented ways, to new, unprecedented challenges and opportunities. There is also a secondary consideration: I will propose that the

[3] Once more: these options are examples useful to make the abstract argument understandable. There is no need to assume that the list provided is exhaustive or universal.

EM is the key that can elucidate why the hard problem has come to exist. The EM evaluates current experience in light of the current motivational state and of previous experience, it then optionally stores the result of this evaluation as additional memories. We have seen that this is not a mysterious computational trickery, but I wish to propose that it is exactly the cause of PE, or more precisely, I will maintain that the EM, when associated with additional and recursive functions, is the fundamental ingredient needed to explain why PE is so puzzling. However, before being able to do so, I need to address the "easy problems", or the side of consciousness that should be studied with a reductionist approach. This is the subject of the next section: I will start by discussing more theoretical or philosophical frameworks and gradually progress toward fully formed scientific theories. The aim is twofold: first, I wish to see if and how existing theories can accommodate the framework sketched so far; second, I will use empirical evidence to extend and refine what is currently just a draft. In this way, I aim to solidify ETC and see if it does account for what is known about consciousness and whether it also provides new answers.

Part two: extending ETC in the light of philosophical and empirical theories.

In this section I will compare ETC with a selection of the most widely accepted theories, starting from the more philosophical ones and proceeding towards models that find their foundations in empirical science. In the process, I will progressively enrich ETC with more details, partially superseding the "agnostic" stance that I have espoused so far (see notes on Part 1).

Before doing so, I need to complete the first draft of the whole model: so far we have encountered three levels: signal collection and transmission (level 1), signal interpretation (level 2) and integration in a unique representation (level 3). Moreover, I've argued that, on top of this, conscious organisms have an Evaluation Module (part of level 4), which is the primary enabler of procedural learning, the ability to learn new, non-stereotyped behaviors. However, once an organism is able to learn in this way, it will benefit from the presence of additional systems that may use the available information in cognitively advanced ways; for simplicity's sake, I will call this collection of modules "executive functions" and place them, along with the EM, inside **Level 4**.

Is ETC yet another Higher Order Thought theory?

From a first, superficial point of view, ETC may be described as a higher order dispositional theory (for review, see Rosenthal 2004): this is because what is represented in level 3 is available to become conscious and the result of the evaluation can be considered to be the "higher order" additional component that is necessary to make a state conscious. Therefore, some perceived information will become conscious, while some other may not, but whatever reaches level 3 is predisposed to become conscious, and becoming conscious implies the creation of a higher order thought (in this case, the result of the evaluation).

However, exploring the implications brings out further questions. Is all the content of the third level evaluated? If not, what is it that makes us evaluate some but not all of the information that is picked up by our senses?

The answers become clear and informative once we postulate that the elements present in level 3 are associated with a "weight" that defines their evaluation priority. In fact, it is well known that some sort of "potential relevance" evaluation is constantly performed along the sensory pathway; attention will therefore

consist in a feedback system that amplifies such relevance, secondarily creating a new "paying attention to" object in the third level representation (see below).

To account for the fact that some available information is never evaluated, so that it never reaches consciousness, one needs to hypothesize that:

1) Information that appears on the third level contains a "priority" measure that is used to regulate the flow into EM.
2) Some fourth-level mechanisms will have the ability to change the priority of third-level elements, as it happens when concentrating on something.
3) Entities are represented in level 3 only when they are justified by sensory input (whether external or internal: level 4 processes may create objects within the self-representation in level 3).

As objects are evaluated according to a schedule that is influenced by their relative weight, some entities represented in level 3 may never become conscious.

Thus, we evaluate something (experience it) because it is pre-emptively marked as likely to be important. If the evaluation confirms and enhances this initial relevance score we may end up concentrating on it. In ETC, the pre-evaluation of potential relevance is always automatic: it is a preliminary, always-on, stochastic evaluation. It is important to point out that there is a growing body of evidence that suggests the existence of a system that regulates what sensory data is going to reach the conscious level, but also indicates that the system itself is very sophisticated and able to respond to surprisingly complex criteria (Brown-Iannuzzi et al., 2014), including personal biases (Balcetis et al., 2012) and the current motivational states.

ETC however does not postulate that a sensation is experienced when the subject is appropriately conscious *of* it (often referred as the Transitivity Principle – TP, see Rosenthal 1997), or that PE happens as and when a higher order thought is associated to a sensory state (unless we decide to classify the result of the evaluation model as a higher order thought). To understand why, it is useful to distinguish between two elements of consciousness: the flow of information that is in the process of being evaluated (coming from level three) and the *self aware* consciousness, when the subject is specifically conscious *of* this information. The first accounts for PE (see conclusion), the second accounts for the ability of *knowing* that we are experiencing a particular sensation. These are two different things that are easily confused, and this confusion explains why different higher order thought (HOT) theories never fully convince (see below).

In ETC, attention is the process of evaluating the details of some element of the third-level model, possibly by gathering more input from levels one and two, but also by generating new (often unconscious) behaviors, such as placing the object of interest in the visual focus. Sometimes attention could require retrieving known (previously memorized) qualities of an entity present on level three. This exemplifies why giving attention to something is a complex, time consuming, and costly activity: as such, it is a state that is useful to represent in the third level. Within level 3, the object that is being examined will acquire more weight, so that data associated with it will be fast-tracked in the evaluation pipeline. At the same time, the self-representation in level 3 will gain the quality of "concentrating on" the object itself. This secondary representation may itself be evaluated, so that the state of *giving attention to something* generates the possibility of becoming conscious in the transitive sense: one can be aware, and may report, that her attention is being focused on something.

Imagine that a subject is examining a ball, the "concentrating on: ball" state will appear, as a property of the self, in level 3. As such, it may itself become subject of attention, subsequently creating the "concentrating on: 'concentrating on ball'" construct. When this is created, the action of creating it makes the fact that we are "concentrating on ball" automatically and inevitably experienced, because creating the *wrapping* object requires the evaluation of the *contained* object. But since evaluating something creates PE, it happens that we are conscious of our own experience, in the *self-aware*, transitive sense, only when we focus on a part of experience itself.

This makes it possible to classify ETC as a HOT theory[4]: it describes the awareness of an experience as an additional representation that could be seen as a HOT. It is worth noting that HOT theories have been subject to extensive criticism (see for example Block 2009) and that they nevertheless continue to remain relevant, in virtue of philosophical (Rosenthal, 2012) and empirical (Lau and Rosenthal, 2011) arguments. Moreover, this account of attention also resembles the theory proposed by Graziano and Kastner (2011).

This explanation of attention and transitive awareness also addresses the concerns of those that argue (see for example Blackmore, 2002) that consciousness is an illusion, that there is no movie being constantly projected inside our brain, and that "consciousness" is only present when we happen to pay attention.

In other words, ETC appears to immediately offer an explanation of the puzzling qualities of the transitive principle: it explains how and why only some parts of experience become fully conscious and at the same time clarifies the role of attention, and why it is intuitively appealing to maintain that a mental state is conscious when the subject is appropriately conscious of that state. At the same time, ETC explains the evanescent qualities of reportable experience, as it provides an explanation of why experience feels complete (every sensation, being represented on the third level, *can* become fully conscious) but in fact it is not: we can focus our attention on only a limited number of representations, and becoming conscious of our focus immediately requires the focus itself to shift.

To conclude this brief philosophical part it is also worth highlighting that ETC can explicitly define the difference between PE and awareness *of* a perception, a distinction that allows us to surpass the supposed incompatibility of first- and higher- order theories. Ned Block has outlined this controversy in the following way (Block, 2011):

> *There are two quite different approaches to the nature of consciousness. One line of thought (that I have favored) emphasizes the notoriously elusive 'what it is like' […]. The second approach is the one that emphasizes that a conscious state is a state one is conscious of.*

My argument here is that both approaches are misled by their own assumptions: in particular, they both try to explain consciousness by *focusing on one part* of the phenomenon (either PE or TP), but neither acknowledges the importance of both. The discussion above does address these distinctions and thus offers a way to unify first- and higher-order theories of consciousness: by accepting that PE is distinct from, and precedes, the self-conscious awareness of PE (TP), ETC accepts the strong criticism against HOT (yes, HOT

[4] A full account of the debate on HOT can easily become very convoluted and would be inappropriate here. It should be sufficient to mention that many question the ability of HOT to provide the explanatory bridge, and/or maintain that HOT theories invariably contain some internal inconsistencies.

theories do not, in themselves provide the explanatory bridge) but does not need to negate the importance of the Transitivity Principle.

A Cartesian Theatre watched by a mere algorithm

Since the third level holds a representation of the world that is constantly updated with sensory input, it does resemble a theatre, and would therefore be vulnerable to Dennett's critique (Dennett, 1991). The theatre metaphor can of course be traced back to Descartes, is the foundation of dualism, and clearly, not the subject of ETC. However, Dennett argues that theories of consciousness can frequently be seen as equally misguided by the intuitive appeal of the theatre metaphor: if there is a stage where consciousness is enacted, there necessarily is an internal observer (or homunculus), which would shift the problem of consciousness inside the "internal observer" and therefore explain exactly nothing. However, besides level 3, ETC postulates two other elements, the EM that probes the significance of elements, and the mechanism of attention. Hence, the third level can be seen as a Cartesian Theatre, but there is no central homunculus watching. Instead, there is a spotlight that evaluates only one or few of the elements present on the stage, and does so in computationally delineable ways (i.e. algorithmically). This evaluation process can generate fully conscious and reportable states by means of secondary representations that only happen when we are actively trying to notice what it is that we are experiencing. In this sense, ETC is suitable to resist Dennett's demolition of the theatre metaphor, but still accounts for the introspective plausibility of the very same idea.

At this point one could argue that the evaluation process is in fact some kind of homunculus, and that its powers of evaluation are too vast and vaguely defined. I have explained in part one why this isn't the case: the EM is a pattern-matching engine, one that could evolve from simpler evolutionary ancestors and that, at the bare minimum, is used to classify sensory information as worth remembering, as something that requires more attention or as trivial data that can be ignored. In other words, looking at ETC in computational terms, it can be said that ETC provides the minimum requirements for the creation of consciousness: I will briefly discuss this approach in the conclusion, along with some important added cautionary remarks.

The role of qualia

It is impossible to study consciousness in philosophical terms without encountering the concept of *qualia*, defined as, on first approximation, "*the raw qualities of feelings*". In the philosophical tradition, qualia are puzzling because they are supposed to have some extraordinary characteristics: traditionally, qualia are irrevocably *ineffable* (can't be fully described in words: how can you describe the difference between yellow and green to a blind person?), *intrinsic* (they are indivisible properties of what is being experienced, they can't be reduced to more fundamental components), *private* (since they are ineffable and intrinsic, they can't be communicated) and directly or immediately *apprehensible in consciousness*. As such, qualia are the subject of the hard problem, they represent the supposedly inexplicable contents of PE, and therefore can only be addressed in philosophical arguments.

Dennett has famously and vehemently attacked the very idea of using the concept, as well as the word "qualia" itself in *Quining Qualia* (1988) where he convincingly argues that PE can't possibly have any one of the qualities listed above. According to Dennett (and ETC fully supports this view), qualia are neither ineffable, intrinsic, private nor directly experienced. Therefore, Dennett advises to avoid using the word *qualia*, in order to stop propagating powerful but mistaken intuitions. I consider myself convinced, and have avoided using these words so far as a result. However, I will now need to disregard Dennett's suggestion and

introduce a discussion on qualia, but a different kind of qualia: the qualia I will discuss are indeed the raw qualities of feeling, but they are approached, defined and discussed in strict neurological terms; as such, they do not show any mysterious or inexplicable quality and can be studied empirically. This long premise was necessary because the work I wish to discuss below does use the word qualia, making it impossible to avoid using the contentious term. Moreover, I need to explicitly warn the reader: although the term quale is loaded with a heavy burden of philosophical connotations, the discussion below starts with new, empirically derived definitions, and just happens to apply them to an old, already in-use term.

The model of qualia I wish to discuss was developed by Ramachandran and Hirstein (R&H, 1998) in their paper "Three laws of qualia: What neurology tells us about the biological functions of consciousness". Their definition is:

> *Qualia are the 'raw feels' of conscious experience: the painfulness of pain, the redness of red. Qualia give human conscious experience the particular character that it has.*

They conclude that all qualia have three characteristics:

1. Irrevocability on the input side
2. Flexibility on the output side
3. Qualia-based representations are held in an intermediate buffer, in 'working memory'.

Their paper contains a long, detailed and convincing collection of real-world, observable examples that are useful to demonstrate that this sort of qualia do exist, and that they indeed follow the three rules above. Therefore, I consider it as the first empirical confirmation of what I have been speculating so far: to see why, I will explicitly match the elements of ETC with R&H's theory of qualia.

1) The existence of a unified stage (level 3) presupposes the need of encoding the result of perception and automatic signal processing (levels 1 and 2). R&H's qualia are nothing more than the different *fundamental symbols* as defined by the required *encoding*.
2) R&H's working memory is in fact level 3 itself.
3) The executive and complex processes that happen in level 4 ensure the "flexibility of output". Even the first operation performed on level 4 (the evaluation) is informed by existing memories, and therefore is inevitably non repeatable (assuming that the memory systems are in working conditions, each evaluation can influence the evaluations that follow).

However, ETC offers two significant additions: it contains more complete explanations of the evolutionary origin of R&H's qualia, and includes a new vision on the problem of the self.

R&H assert:

> *The single most important principle underlying the mechanisms of perception and conscious experience [is] that they may have evolved exclusively for extracting statistical regularities from the natural world.*

I don't think that this statement tells the full story. The zombie-like lizard of my first example already has all the machinery necessary to extract *some* statistical regularities from the world. In ETC, this task is necessary to build a model of the world that will inform the behavioral responses of a zombie-like organism, but does not require the experience phenomenon: zombie-like organisms lack experience because they do not need

to provide a conceptual bridge between the complex internal dispositions and what happens on the outside. In other words, I am suggesting that proto-qualia are already present in non-conscious organisms, and that they are entities used to describe the world also in their case. However, proto-qualia do not associate with experience because non-conscious organisms do not have the ability to learn entirely new strategies and therefore proto-qualia do not strictly obey to the second law of qualia (the output is somewhat fixed). The upside of this little disagreement is that ETC remains fully supported by all of R&H arguments, but it does offer more explanatory power, thanks to the additional details and distinctions included.

In the last part of their paper, R&H address the problem of the self, and argue that this is what uses the qualia-based representations. They also hypothesize that:

> *[this self is] a process involved in connecting motivation and emotion with the choice of actions to perform, […] very much the sort of thing which the self was traditionally supposed to do.*

It may also be seen as a collection of limbic executive processes, or a gateway between perceptions and a wider set of executive processes. In ETC, this "self" corresponds exactly with the evaluation process, while the wider set of limbic processes is what I have referred to as the other modules of level 4. Again, both the similarities and the differences between ETC and R&H's theory can be interpreted as a confirmation of the validity of the first, and an extension of the latter. ETC postulates the existence of a self-representation (available for evaluation and therefore experienceable) that contains the representations of internal states (including posture, physiological states, desires and reportable states of the executive processes). Conveniently, such model was already present in non-conscious organisms, in order to allow the existence of complex but fixed behavioral strategies, and was therefore available for extension, following a typical evolutionary paradigm.

We can therefore distinguish *two separate entities that may both be called the "self"*. The **first** is the collection of executive processes that form level 4; the **second** is the representation of the "self" held on level 3.

The executive processes are the self that understands the world and that makes decisions. On the other hand, the "self-representation" is what justifies the *perception* of the self: it is itself available for evaluation. Furthermore, ETC also explains why some executive processes are conscious: they are those that have a relevant effect on the third level (those that can and should be associated with causal evaluations), so that "paying attention to", "trying to remember something" and "making a decision" states can be represented on level 3, and may therefore become conscious. However, there is no evolutionary need to build and maintain a detailed representation of *how* these states function in detail. Therefore we may be aware that we are making a decision, but we'll never know exactly how we are reaching it. This ultimately justifies our widely accepted perception of free will as well as the perception of it as of something akin to a black-box. Our sense of agency is the perception we experience when the "making a decision" state is present on level 3 and is itself evaluated. It feels somewhat mysterious because we possess no (R&H's) qualia that can build a detailed description of its inner workings; it is, in fact, a quale itself.

Global Workplace and Dynamic Core theories

So far, ETC has provided an evolutionary explanation that accounts for the selective advantage associated with consciousness. However, it would remain a fruitless intellectual effort if it could not be validated by current empirical knowledge. The volume of experimental data on the subject of consciousness is unsurprisingly vast, and growing at an accelerating pace (for an up to date review of the recent empirical developments, see Boly et al., 2013); unfortunately, space restrictions make it impossible to discuss a representative selection of the experimental data in light of ETC. Instead, I will look at the theoretical framework that appears to have inspired most of the recent research: the Global Workplace (GW - Baars 1988), and then compare my point of view with the latest theoretical adaptation of GW, an effort that tries to unify the GW and Dynamic Core (DC – Edelman and Tononi 2000) hypotheses. By showing how ETC can be seen as a refinement of the latest efforts, I hope to confirm the validity of my hypothesis and consequently inspire new research.

GW theory was built on the basis of a large and well-documented body of neurological and experimental neurophysiological knowledge, and it is practical enough to provide a range of working hypotheses that can be tested in scientifically sound ways. My understanding is that it offers an explanation of the *function* of consciousness that is equivalent to what I am suggesting: consciousness is the mechanism that allows the maximum degree of behavioral adaptability. Where a strong disagreement may lie, is in the explanation of why and how the PE happens. According to its original formulation, something becomes conscious when and *because* it is broadcast across the brain to a wide range of different brain regions. This point of view has very solid experimental data that support it (Edelman et al., 2011), to the point that it is now impossible to deny that conscious processing is *associated* with a widespread activation of numerous brain regions. However this explanation does not convincingly address (or disprove the existence of) the hard problem. There are several distinct computational functions that are necessary to create both experience and consciousness, and reducing them all to a generic broadcast doesn't seem to account for the details.

To explain why this is so, I will use a first person example, with apologies for the trivial nature of the "experimental settings". As it happens, I like indulging in playing Windows Solitaire. After years of practice, I have refined a general strategy that allows me to play the game while using procedural automatisms only: I am able to play while talking on the phone, and to choose and perform the next move without any conscious thought. Computationally, this is a very complex task: I evaluate the visible cards, identify the significant information and apply a set of rigid rules to determine the next move. This is done in far less than a second, allowing me to conclude that when I play at full speed, I only use what is usually referred as procedural memory. When a low score motivates me, I can switch to more complex strategies that cannot be performed in the absence of conscious thought. This second approach requires to consciously remember cards encountered before and to perform little simulations about the consequences of different moves. In other words, it requires active attention and is incompatible with other complex cognitive functions.

This observation highlights the already discussed distinction between perception (PE) and self-awareness (TP), or between conscious and unconscious perception. There is no doubt that when playing in the mindless mode I perceive the cards in play and even assign a meaning to each one of them. However, I fail to register any relevant information: it is utterly impossible for me to reconstruct what I did or what I saw just a few seconds before the last move. In other words, I can perceive and process symbolic information, even without explicit awareness.

Taken in the context of ETC, I can argue that during mindless playing the evaluation process is fully operational: I do extract a meaning out of the perceptual input, and in fact, I do experience them, as it is possible (with some effort) to recollect the immediately past experience (but not what happened before that). However, I am not conscious of my perceptions in a fully formed sense: PE is happening, but without transitive awareness. On the other hand, when using the fully-engaged approach, I will become able to explain or question my decisions, register errors and adjust my strategy (PE still happens, with added transitive awareness)

I can therefore affirm that the mindless mode is largely analogous to the "hide!" strategy shown by the lizard discussed in part 1: I am giving a meaning to my sensory input, use a goal (have trivial fun) to select and apply a static computational strategy that drives my stereotyped behavioral response. The only difference between the lizard and me is that the computational strategy used to hide is innate, while the one I use to play solitaire was built via conscious effort. On the other hand, when I am playing in the fully engaged mode, I am consciously using the results of the first automatic evaluation to actively make projects and eventually refine the automatically executable set of rules.

ETC proposes the existence of an EM that consumes internal representations in real-time, applies conceptual, symbolic meaning to them, and routes the result either to automatic procedures (procedural memory) or to higher thought systems (when using the "fully engaged" mode). In the first route, experience happens (PE), but does not necessarily reach full consciousness (TP). When the latter route is followed, various degrees of consciousness may be achieved, from transitive awareness and all the way up to verbal reportability.

In other words, the Solitaire example highlights the subtle distinction between perception and full (transitive) consciousness, but it also leads us to clearly highlight the main shortcomings of the GW theory.

While playing mindlessly, I am actually performing quite remarkable tasks: I'm isolating the content of the solitaire window from the background, then the single cards, identify the relationships between cards, and route the result towards a procedural system. All this happens automatically, and without explicit consciousness. To do all this, vast amounts of information need to be accessed and processed according to a complex set of rules. Once this is done, motor signals need to be produced and coordinated, all in less than a second. In other words, the coordinated activity postulated by GW does indeed happen, but it is in part necessary to generate perception, and in part to consume the evaluation results. The logical consequence is that the GW theory risks confusing the empirically observable neural correlates of PE (not even the neural correlates of full consciousness) with the phenomenon itself: even a task that involves only procedural automatisms requires to evaluate, access and process a lot of information. Thus, it will generate plenty of neural activity that correlates with experience but is involved with many other functions. My objection to classic GW therefore is: on one side, claiming that consciousness is generated by a broadcast does not address the problem of PE; on the other, I claim that a global (or wide) broadcast does indeed need to happen, but it is caused by the systems that sustain consciousness, it is not the cause of consciousness.

The second weakness of the GW theory is experimental: it is difficult to design and validate experimental paradigms that allow the exclusion of unwanted mechanisms in order to highlight the neural activity responsible for the single parts of the consciousness system. First of all, to know if a stimulus was perceived, some kind of output needs to be recorded. Even if the subject only needs to press a button in response of a

given clue, she will still activate a range of motor areas, and we have seen that when properly trained, this happens well below the temporal resolution of even the most modern noninvasive neuroimaging techniques (Ugurbil et al., 2013).

Furthermore, if one wants to discriminate between experience and full (transitive) consciousness, it is necessary to investigate subjects that can decide not to be conscious of what they are doing, have them properly trained, and still find a way of isolate the activity that correlates to perception itself, from all the further processing.

Because of all these difficulties, most of the experimental settings reported in the relevant literature seem to have been designed by scientist that were prepared to overlook the weaknesses explained above, and have instead concentrated on isolating and classifying the neural correlates of consciousness in a very broad sense. I must add that this limitation is also the strength of the GW theory: by providing a somewhat simplified theoretical framework, it enabled the design of a remarkable amount of experiments, and is therefore responsible for the accumulation of very valuable knowledge. However, it seems clear that the scientific community is becoming aware of the limitations of the theory, and is already striving to find new ways to overcome them (see for example Sergent and Naccache, 2012).

In fact, Baars himself has explicitly addressed these shortcomings in Edelman et al. (2011), where they propose an integrated model that I wish to discuss further. In their premises, Edelman and colleagues start their discussions by pointing out that:

> *At the outset, it is important to distinguish primary sensorimotor consciousness from higher-order consciousness […]. Primary consciousness occurs in animals lacking any linguistic capabilities, and is an essential process even in humans.*

> *[…] In the vertebrates, consciousness is a dynamic, integrated, multimodal mental process entailed by physical events occurring in the forebrain. Plausibly, the neural structures and mechanisms underlying consciousness were selected during evolution because they allowed animals to plan and prepare for future eventualities.*

Furthermore, they integrate GW and DC theories, by postulating that the global activation depends on what happens in the DC.

> *Dynamic Core activity […] generates conceptual content by integrating both perceptually driven and motor signals with stored memories. The fundamental tie to world signals in turn provides a basis for intentionality.*

These views fully support ETC: the DC is functionally equivalent to what I have called the Evaluation Process; it provokes the global activation and is responsible for the creation of conceptual content, using perceptual information as input. It is also the fundamental tie between purpose, the outside world and behavioral output. In this sense, ETC is fully supported by the arguments of Edelman et al., and it is further extended by their proposed physiological explanations of the underlying neural networks. It is also complementary, because it engages in the deeper conceptual explanation of the exact computational roles and functions of the DC. However, my views do not match the theory exposed in their paper when they try to engage with two philosophical concepts: the self and the hard problem.

As seen above, ETC accepts the existence of two separate selves: one is the consumer of perceptual input, the pool of processes that perform the first evaluation, along with the superstructures that consume the evaluation results. The second "self" is the internal representation of the self, a representation that is itself *perceivable*. Edelman et al. seem to acknowledge the existence of this dichotomy, but fail to make it explicit, and instead feed the source of confusion. They ask, "Exactly 'who' experiences qualia in a given body? Or, more succinctly, what is the self?", and instead of solving the problem by acknowledging the existence of two separate selves, they propose the existence of the second self alone, the one that is *perceivable* and available to the DC evaluation.

Furthermore, they dismiss the hard problem by noting that even if the qualia sensations seem to be causal of our behavior, in fact they are not: they are the internal perception of the real causal entity, the neuronal mechanism. From an ideal, purely objective perspective, this may be seen as true, and I agree that if the aim is to understand the mechanistic nature of our brain, the hard problem may be brushed aside. However, understanding how to link scientific facts with subjective experience is equally important. Hence, instead of dismissing subjective experience as irrelevant, it would be more interesting to combine the following observations: the physical brain is the sole responsible of our behavior, and this needs to be accounted by means of physical, molecular, neuronal and systemic explanations; but in the same way, our thoughts and beliefs are the basis of our choices, and hence of our behavior. In other words, there is indeed an explanatory bridge to build, and I see no reason to believe that it is impossible to do so. It is, on the contrary, quite easy: beliefs and perceptions are not epiphenomena that somehow arise from neural activity; they are in fact the neural activity itself, as observed from the subjective point of view. The distinction is the direct consequence of two epistemologically separate reference systems. One is subjective experience, which will inevitably produce a different account of the phenomenon of experience when compared with detailed, third-person, scientific descriptions. We have discussed how qualia form the basis of a descriptive language that is used to subjectively evaluate the meaning of what is perceived; it is therefore also true that the neural activity that generates and then evaluates qualia is one and the same as subjective experience. When my brain generates the neuro-physiological representation of the two of diamonds, and then processes this representation to discriminate its meaning, I will see/perceive the card, because that is how the same neural activity is registered from within. The representation however is one and one alone: from the objective point of view, it will one day be described as a collection of connected neurons, action potentials, and who knows what else. From within, it is just a playing card that can be described in terms of symbols, colors, patterns and materials. The two descriptions seem to apply to separate entities only because of the irreconcilable differences between the two distinct languages used to build them, but they in fact refer to the same real entity, as seen by two different perspectives and described by separated symbolic languages.

Information integration theory and predictive coding

I will return to these concepts, but before doing so, I wish to quickly discuss two other scientific theories of consciousness: Tononi's Information Integration Theory (IIT - Tononi 2012, Oizumi et al. 2014), and predictive coding (PC). Tononi's IIT is only partially incompatible with what I'm proposing here: in fact, Tononi's point of view can be accepted as a credible hypothesis as long as his claims of having produced a full model are disregarded (for some hard hitting criticism, see Schwitzgebel 2014 and Aaronson 2014). In my opinion, the mathematical core of IIT is a very promising model of how information may be represented and interrogated when it reaches level 3: it provides an example of what the language that is used to describe the level 3

representations may look like. IIT shows how representations may be built in a unified manner, integrating different sensory information and already developed concepts. This is an outstanding achievement, but it does not mean that any system that uses a similar integrative language is necessarily conscious: it does not in itself explain PE. Furthermore, IIT is almost unapologetically panpsychist (see Tononi and Koch 2015 for the apologetic part), and does little to explain how learning may work.

Learning itself is in fact at the heart of the latest and exciting neuroscience hypotheses: the Bayesian brain and PC (see Clark, 2013). Because of space restrictions, I am not able to discuss in detail how ETC relates to PC, but I do wish to include a few remarks.

PC can be seen as a computational theory that aims to explain how brains may generate and apply useful meanings to perceptions (amongst other things), and postulates a hierarchical generative model that operates at increasingly abstract levels. ETC can be integrated within the PC framework by observing that:

1) All steps from level 1 to EM are essentially a pattern-matching exercise; PC may be used to explain how this matching is done, and how it can be refined as new inputs are processed. PC postulates that all levels of processing are able to learn, something that does contribute to ETC.
2) However, PC may be used to create a continuous route from perception to action, but doing so is problematic, because it seems to ignore the role of agency, and could arguably all happen in the dark (without PE). ETC offers a neat solution to this problem: it explains how conscious organisms may achieve intentionality while operating within the PC paradigm. This happens because the EM operates where two distinct pathways converge: sensory systems are responsible for generating the model of the world, while interoception generates the model of the self. Hence, the EM can be seen as the PC engine that is responsible for modelling the relation between internal states (including motivational ones) and the outside world. PC, in other words, provides an elegant formalization of what I have so far described as “giving perceptions a meaning”.
3) In other words, ETC addresses the problem of how PC may sustain volition: it is generated by modelling the relationship between motivational states and sensory information; volition happens at the intersection of two very separate information pathways, and can be seen as the point where a discontinuity is introduced between sensory information and behavior.

Conclusions

After proposing an original first draft, constructed using evolutionary principles in part 1, I have used existing philosophical and empirical theories to check the solidity of the fundamental architecture and to add more detailed descriptions. Only after doing so it has become possible to describe ETC in full. Before discussing its testability and how it may aid future research, it is useful to summarize the whole theory, providing some explicit statements that may be tested individually.

- ETC states that sensory information, proprioception and representation of internal states all flow into a model of the world and self, what I have called level 3.
- Every element present in level 3 includes a relative weight that is used to prioritize further processing.
- The first subsequent step is to evaluate the significance of elements in level 3, in order to identify (at least) what perceptions may be worth remembering, what requires immediate action, and what should be given further attention.
- ETC argues that this discriminating ability is necessary to allow organisms to learn new adaptive behaviors.

- The evaluation can generate new representations in the "self" side of level 3, for example by signaling that the subject is focusing its attention on the just-evaluated perception.
- PE is generated by the evaluation, but full, transitive, *self-aware* experience is only obtained when the result of the evaluation (above) is itself recursively evaluated (a process mediated by attention).
- ETC therefore distinguishes two pairs of easily confused elements:
    - The first pair is formed by PE (the evaluation itself) and awareness of perception (TP), which is obtained by recursive evaluation.
    - The second pair regards the self. ETC distinguishes between a model of the self (and of its internal states) that can be experienced, and a computational self, responsible for generating PE, transitive awareness and behaviors.

I short, ETC claims that PE is the necessary consequence of the ability to learn, which in turn requires the ability to store memories of experience, and to link them with adaptive, genetically encoded, purposes. Thus, ETC contains a philosophical answer to the supposedly hard problem of consciousness, or at least, it provides the foundation necessary to build the explanatory bridge between mere computations and PE itself. After describing ETC in full, it is finally possible to defend this claim.

One way of describing the hard problem is as follows: *how does it happen that certain particular computations* (e.g. transforming and manipulating a given signal)*, generate inner feelings?* After explaining the functions that operate in ETC, I will now reverse this question, and ask: *how could the functions of ETC not generate inner feelings?*

Let us follow the computational route, and see what happens when an ETC system is actively trying to understand its inner workings. At first, the result of sensory stimuli reaches the EM, which has the task of answering the following question: does this perception mean something to me (the *experiencing* subject)? In this way, the EM applies a very basic form of "meaning" to perceptions, nothing more than the classification outlined at the end of part 1 (e.g. worth remembering, requiring immediate action, attention, etc) – I will use the word "meaning" in this sense, unless otherwise specified. Note also that "giving a meaning" implies that perceptions are classified according to pre-existing categories that the evaluation system is already considering useful (meaningful to the perceiving subject): some of them will be innate, but the vast majority will be built as the result of previous classifications. Crucially, to perform this classification, the EM has access to a constantly updated model of the self and to memories of previous experiences; thus, a high-level, philosophical way of describing its task is to say that the *EM generates a point of view*. It applies meaning to perception, but we have seen how ETC allows us to define this task in strictly computational terms. In doing so, it evaluates a symbolic description of reality (built using the fundamental symbols that are R&H's kind of qualia) and adds more information to it: it enriches incoming signals with their subjective meaning.

However, the exploration of what happens while I play Solitaire, along with the insights provided by the analysis of the TP, allow us to conclude that this direct evaluation is not the end of the story. When a subject is actively trying to understand how its own perceptions work, it will need to re-evaluate the evaluation. The first part of part 2 describes the process: the *result* of a first evaluation is stored in working memory (in particular, as a new representation inside the model of the self) and this second object is itself re-evaluated. This means that the original symbolic representation, now enriched with the newly added meaning, is once more "looked at", providing an input that may be described as follows: "I perceived a red apple, and it's good to know, because I like apples". From within the EM, the redness of the apple is an ineffable, indivisible quality of the original perception, now enriched with subjective value ("I like apples"). However, re-evaluating the result of a perception finally makes the subject aware of the perception: the redness of the

apple is now conscious, because the new evaluation is re-accessing it (potentially generating a third object, describable as "I know I have perceived a red apple"). The nature of this redness is still mysterious to the EM itself because "red" is symbolically encoded and the details of how this description works (the code used) are themselves inaccessible to the EM (as this module is performing a mere algorithm). The result of recursively evaluating our perceptions, with the explicit purpose of introspectively understanding what PE is, generates in this way the idea of the older kind of qualia: the ineffable, intrinsic, private and directly experienced ones. When trying to figure out what makes the apple red, all the EM can find is that the apple just has this redness quality, which is its own perception of perception, and has to be mysterious, because all the internal computations rely on a descriptive code that can never in itself be the input of another evaluation.

Recursive re-evaluations of perceptions can't be performed *in the dark* because they are the attempt of applying a subjective meaning to already evaluated perceptions, and this exposes the symbolic nature of the evaluated information, showing that the fundamental symbols are, from within, indivisible and impermeable to further introspection (e.g. they appear to the experiencing subject as the mysterious, *philosophical* qualia that just *are*).

ETC predicts thus that when a computational system that follows ETC tries to introspectively analyze the nature of its own computations, it would find the mystery of PE, or, if you prefer, it would experience the puzzling qualities of philosophical qualia.

It is also worth noting that the argument I have just made, in algorithmic or computational terms, is effectively proposed, in an alternative way, by Loorits (2014); his central claim, which uses the theory proposed by Crick and Koch (1998), is nevertheless built on philosophical grounds, and is explicitly applicable to alternative empirical theories of consciousness. In particular, Loorits explains why:

> *Qualia are highly complex and perfectly public structural-relational properties of some cognitive systems, even though those systems themselves perceive them as monadic and private.*

Which is precisely my point: qualia, when exclusively analyzed via introspection, show puzzling properties, but these are the direct consequence of the intrinsic limitations of what consciousness does, and are, in fact, objectively misleading.

Therefore, I argue that the thought experiment performed above explains the existence of the hard problem, it shows how evolutionary pressures made it emerge. But does it also provide the explanatory bridge? No, it does not provide a full bridge, but it does point to a promising direction: it tells us that understanding how information is both encoded and evaluated will indeed provide two fundamental pieces of the puzzle. It also tells us that experience is the predictable consequence of selective pressures. In this view, the hard problem is the result of behavioral plasticity, paired with enough computational power. ETC predicts that an ETC system that is equipped with enough computational resources, and is at the same time animated by the purpose of understanding reality, will eventually face the problem of PE. Thus, ETC shows that it is possible to build a bridge between a third-party, computational account of how the brain functions and our direct experience of mental activity.

If ETC turns out to be accurate enough, it follows that the perceived hardness of the Hard Problem is the natural consequence of our own computational architecture. In the absence of detailed physiological

descriptions, ETC cannot be able to fully bridge the explanatory gap, but it does explain why the gap exists, and more importantly, it provides reasons to believe that a full functional account is indeed necessary to build the remaining part of the bridge.

In other words, ETC is a promising theory, able to exploit, illuminate and redefine one of the most impervious problems known to both science and philosophy. Still, in order to do so, it needs to be validated. This need is usually translates to the following question: can ETC be empirically falsified?

One, seemingly obvious, prediction of ETC can be described as follows: conscious organisms will be unable to learn new behaviors if and when the relevant information is registered but does not reach full, transitive consciousness. Testing by experiment this hypothesis should not be difficult: it can be done following established stimulus masking protocols (Lu and Dosher, 2013). However, on the face of it, the prediction seems so obviously true that many may object that it provides little or no significant confirmation. This may be the case, but I wish to argue that this objections rests on unwarranted assumptions, or, more precisely, I wish to point out that the intuitive obviousness of this prediction is actually questioned by recent evidence. The work by Balcetis (2012) and Brown-Iannuzzi (2013) tells us that what may reach self-aware consciousness is heavily influenced by our current motivational state, as well as by our unconscious biases. This in turn implies that stimuli are already filtered and categorized in very complex and dynamic ways well before they reach the conscious level. The consequence is that there is no *a priori* reason to expect that these evaluations cannot be used to learn how to navigate the world. In this context, ETC provides an *a posteriori* reason: it explicitly hypothesizes the existence of computational modules that are specifically dedicated to learning, and that their function generates PE. Therefore, ETC predicts that PE is a necessary condition to learn new behaviors, even if the current evidence does not imply this otherwise intuitive conjecture.

The above all-important prediction is the one make it or break it verifiable claim that I wish to make: as explained in the introduction, ETC is purposely highly abstract, precisely because it attempts to become useful in evaluating and bridging together pre-existing theories. Thus ETC needs to introduce the smaller possible number of core hypotheses, and leave as many details as possible available for subsequent specification. I will now briefly hint how this "subsequent specification" may be done, linking to the recommendations made by Adolphs (see the Introduction): to progress, cognitive neuroscience needs to focus on computations and take a comparative approach.

The computational application of ETC should be quite clear: given a fully described neural circuitry (thanks to the current large scale mapping projects, there is hope that such comprehensive maps will proliferate in the near future), one could directly ask the questions "is this particular network able to perform a pattern-matching function similar to the one proposed for the EM?", or "are these re-entrant connections suitable to provide the kind of recursion that ETC links to attention?". How to verify all these second-order hypotheses would naturally depend on the combination of structure/function pairs, the mid-level theory adopted, and the feasibility of different experimental approaches; as shown in the discussion above, ETC could be used to augment and refine theories as separate as DC-GW and Bayesian approaches.

Answering this sort of questions would then permit to produce a range of subsequent hypotheses on the detailed functions of structures such as the Cortico-Thalamic system, and/or to hypothesize alternative function/structure pairing, based on what algorithmic capacities are shown by the corresponding networks

(mid-level simulations may guide this pairing process). In this way, ETC provides a crucial element in the sort of framework that is necessary to compare the proposed functions of different networks, and/or compare alternative interpretations of what kind of computations are performed by a given network. At the same time, it would allow to explore, from a radically different perspective, the study of consciousness in non-human, and non-mammalian species (see for example Butler 2007). When studying non-human subjects, the overarching hypothesis is that consciousness tightly correlates with higher order cognitive abilities; this intuitive premise is directly supported by ETC, as it provides a theoretical explanation of why learning abilities (at the core of all higher order cognition) may be necessarily tied to consciousness. Furthermore, the work of Butler and others is showing us the evolutionary analogies between the mammalian neocortex and avian pallium: this is another confirmation of the basic foundations of ECT as it is indirect evidence that the evolutionary pathway hypothesized in part 1 is in fact likely to occur.

In other words, ETC finds its support on evolutionary, philosophical and empirical ground and is easily and directly falsifiable, making it a sound, if untested, scientific hypothesis. Moreover, if verified, ETC promises to provide a much-needed conceptual guide to start linking together the different explanatory theories that are associated and tightly linked-with the different levels of enquiry that converge in the whole field of Cognitive Neuroscience.

Ethical considerations

Should ETC be validated and widely accepted, what may the consequences on animal experimentation, and animal exploitation for other ends, be? This is hard to say, but it is easy to predict that the implications would be wide and far-reaching.

Another ethical problem generated by ETC comes from the algorithmic description: it can be read as a high-level guide of how to build a consciousness simulation or a conscious machine. This is a deeply troubling perspective: if it were possible to design and implement an artificial system that is demonstrably organized in a way that reflects ETC, should we consider it conscious? Would it be ethical to tinker with it, and even to arbitrarily turn it on and off at will? Furthermore, I wish to highlight that the current model presents a *clear and very concrete danger*: if an artificially conscious system is created, it will, by definition, have its own, hardwired set of purposes. This may be safe if they are limited to things like winning at chess or sustaining a conversation with human beings, but would become very disturbing if any artificial consciousness had elements that are even remotely connected with self-replication: evolutionary theory tells us that if a learning machine is actively trying to self-replicate, it may eventually succeed, and the risk that it will escape human control is almost too great to even contemplate. Therefore I wish to explicitly say that trying to generate an artificial consciousness following the ETC guidelines may be feasible, and it does certainly look like a thrilling perspective, but it should be approached with extreme caution and integrated with all conceivable internal (built-in purposes that are deemed to be intrinsically safe) and external containment facilities. It may seem inappropriate to linger on such precautionary and arguably premature warnings, but I do not believe it is: if the present theory can enable the creation of an artificial consciousness, such concerns will inevitably become relevant at the flick of a switch.

Acknowledgements

I am indebted to four anonymous peer reviewers and one editor (Dr. Daniel Pollen) already: each one of them has contributed to this manuscript in significant, sometimes challenging, but always productive ways. I

also wish to send a very special "thank you" to the whole of the Social Science Research Unit (SSRU – UCL Institute of Education) for all of their support and for being such an extraordinary group of people: my gratitude goes in particular to Professors David Gough, James Thomas and Ann Oakley, as well as Ginny Brunton and Jeff Brunton. I also wish to thank Prof. Susan Blackmore and Peter Hankins for invaluable advice, generous support and fruitful discussions. Finally, I should thank Prof. Lucia Boldrini for her infinite patience, constant support and for her transformative influence.

References

Aaronson, S. (2014), Giulio Tononi and Me: A Phi-nal Exchange, from *Shtetl-Optimized The Blog of Scott Aaronson*, accessed 22/08/2015. http://www.scottaaronson.com/blog/?p=1823.

Aaronson, S. (05/30/2014 2014). Giulio Tononi and Me: A Phi-nal Exchange. *Shtetl-Optimized The Blog of Scott Aaronson* [Online]. Available from: http://www.scottaaronson.com/blog/?p=1823 [Accessed 30/05/2014 2015].

Adolphs, R. (2015). The unsolved problems of neuroscience. *Trends Cogn Sci* 19**,** 173-175. doi: 10.1016/j.tics.2015.01.007.

Baars, B.J. (1988). *A cognitive theory of consciousness.* Cambridge: Cambridge University Press.

Balcetis, E., Dunning, D., and Granot, Y. (2012). Subjective value determines initial dominance in binocular rivalry. *Journal of Experimental Social Psychology* 48**,** 122-129. doi: http://dx.doi.org/10.1016/j.jesp.2011.08.009.

Blackmore, S. (2002). There is no stream of consciousness. *Journal of Consciousness Studies* 9**,** 17-28.

Block, N. (2009). "Comparing the Major Theories of Consciousness," in *The cognitive neurosciences,* ed. M.S. Gazzaniga. 4th ed. ed (Cambridge, Mass. ; London: MIT Press), 1111-1122.

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis* 71**,** 419-431. doi: 10.1093/analys/anr037.

Boly, M., Seth, A.K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., Edelman, D.B., and Tsuchiya, N. (2013). Consciousness in humans and non-human animals: recent advances and future directions. *Front Psychol* 4**,** 625. doi: 10.3389/fpsyg.2013.00625.

Brown-Iannuzzi, J.L., Hoffman, K.M., Payne, B.K., and Trawalter, S. (2014). The invisible man: interpersonal goals moderate inattentional blindness to African Americans. *J Exp Psychol Gen* 143**,** 33-37. doi: 10.1037/a0031407.

Butler, A.B. (2008). Evolution of brains, cognition, and consciousness. *Brain Research Bulletin* 75**,** 442-449. doi: http://dx.doi.org/10.1016/j.brainresbull.2007.10.017.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 36**,** 181-204. doi: 10.1017/S0140525X12000477.

Crick, F., and Koch, C. (1998). Consciousness and neuroscience. *Cereb Cortex* 8**,** 97-107.

Damper, R.I. (2000). Editorial for the special issue on 'Emergent properties of complex systems' - Emergence and levels of abstraction. *International Journal of Systems Science* 31**,** 811-818. doi: Doi 10.1080/002077200406543.

Dennett, D.C. (1988). "Quining qualia," in *Consciousness in contemporary science,* eds. A.J. Marcel & E. Bisiach. (Oxford: Oxford University Press), 42-47.

Dennett, D.C. (1991). *Consciousness explained.* London: Allen Lane The Penguin Press.

Edelman, G.M., Gally, J.A., and Baars, B.J. (2011). Biology of consciousness. *Front Psychol* 2**,** 4. doi: 10.3389/fpsyg.2011.00004.

Edelman, G.M., and Tononi, G. (2000). *A universe of consciousness : how matter becomes imagination.* New York, NY: Basic Books.

Graziano, M.S., and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cogn Neurosci* 2**,** 98-113. doi: 10.1080/17588928.2011.565121.

Kirk, R. 2012. Zombies. in *The Stanford Encyclopedia of Philosophy* [Online]*,* ed. E.N. Zalta. Available: http://plato.stanford.edu/archives/sum2012/entries/zombies/ [Accessed 13 Oct 2013].

Kurahashi, T., and Menini, A. (1997). Mechanism of odorant adaptation in the olfactory receptor cell. *Nature* 385**,** 725-729. doi: 10.1038/385725a0.

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* 15**,** 365-373. doi: 10.1016/j.tics.2011.05.009.

Levick, W.R. (1967). Receptive fields and trigger features of ganglion cells in the visual streak of the rabbit's retina. *The Journal of physiology* 188**,** 285-307.

Loorits, K. (2014). Structural qualia: a solution to the hard problem of consciousness. *Front Psychol* 5**,** 237. doi: 10.3389/fpsyg.2014.00237.

Lu, Z.-L.A., and Dosher, B.A. (2013). *Visual psychophysics : from laboratory to theory.*

Marcus, G.F., and Freeman, J.A. (2015). *The future of the brain : essays by the world's leading neuroscientists.* Princeton: Princeton University Press.

Metzinger, T. (2003). *Being no one : the self-model theory of subjectivity.* Cambridge, Mass.: MIT Press.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS computational biology* 10**,** e1003588.

Ramachandran, V., and Hirstein, W. (1997). Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies*.

Rosenthal, D. (1997). "A Theory of Consciousness," in *The nature of consciousness : philosophical debates,* eds. N. Block, O.J. Flanagan & G.V. GüZeldere. (Cambridge, Mass.: MIT Press), 729-753.

Rosenthal, D. (2004). "Varieties of higher-order theory," in *Higher-order theories of consciousness : an anthology,* ed. R.J. Gennaro. (Amsterdam ; Philadelphia, PA: John Benjamins Pub.), 19–44.

Rosenthal, D. (2012). Higher-order awareness, misrepresentation and function. *Philos Trans R Soc Lond B Biol Sci* 367**,** 1424-1438. doi: 10.1098/rstb.2011.0353.

Saito, S., Fukuta, N., Shingai, R., and Tominaga, M. (2011). Evolution of vertebrate transient receptor potential vanilloid 3 channels: opposite temperature sensitivity between mammals and western clawed frogs. *PLoS Genet* 7**,** e1002041. doi: 10.1371/journal.pgen.1002041.

Sawyer, R.K. (2002). Emergence in psychology: Lessons from the history of non-reductionist science. *Human Development* 45**,** 2-28. doi: Doi 10.1159/000048148.

Schwitzgebel, E. (16/07/2014 2014). Tononi's Exclusion Postulate Would Make Consciousness (Nearly) Irrelevant. *The Splintered Mind* [Online]. Available from: http://schwitzsplinters.blogspot.co.uk/2014/07/tononis-exclusion-postulate-would-make.html [Accessed 16/07/2014 2015].

Searle, J. (2013). Theory of mind and Darwin's legacy. *Proc Natl Acad Sci U S A* 110 Suppl 2**,** 10343-10348. doi: 10.1073/pnas.1301214110.

Sergent, C., and Naccache, L. (2012). Imaging neural signatures of consciousness: 'what', 'when', 'where' and 'how' does it work? *Arch Ital Biol* 150**,** 91-106. doi: 10.4449/aib.v150i2.1270.

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch Ital Biol* 150**,** 293-329.

Ugurbil, K., Xu, J., Auerbach, E.J., Moeller, S., Vu, A.T., Duarte-Carvajalino, J.M., Lenglet, C., Wu, X., Schmitter, S., Van De Moortele, P.F., Strupp, J., Sapiro, G., De Martino, F., Wang, D., Harel, N., Garwood, M., Chen, L., Feinberg, D.A., Smith, S.M., Miller, K.L., Sotiropoulos, S.N., Jbabdi, S., Andersson, J.L., Behrens, T.E., Glasser, M.F., Van Essen, D.C., Yacoub, E., and Consortium, W.U.-M.H. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* 80**,** 80-104. doi: 10.1016/j.neuroimage.2013.05.012.

Vecera, S.P., Cosman, J.D., Vatterott, D.B., and Roper, Z.J.J. (2014). The Control of Visual Attention: Toward a Unified Account. *Psychology of Learning and Motivation, Vol 60* 60**,** 303-347. doi: Doi 10.1016/B978-0-12-800090-8.00008-1.

1009