The Integrated Information Theory facing the Hard Problem of Consciousness

Adrien Wael BASILLE
2005031916T

Travail Encadré de Recherche

Dirigé par Pascal Ludwig

M1 Philosophie
Section Philosophie des sciences, de la connaissance et de l'esprit
Sorbonne Université

Juin 2020

Introduction

The Integrated Information Theory (IIT) formulated for the first time in 2004 by the neuroscientist Giulio Tononi, is a theoretical framework aiming to explain consciousness (Tononi, 2004). What is consciousness? In his more recent presentation of the IIT, Tononi starts by giving a straightforward definition:

> *Everybody knows what consciousness is: it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream. Thus, consciousness is synonymous with experience – any experience – of shapes or sounds, thoughts or emotions, about the world or about the self.* (Tononi, 2012)

This definition corresponds to what is usually called *phenomenal consciousness* in contemporary philosophy of mind. Ned Block distinguishes phenomenal consciousness from *access consciousness*, the latter being functionally defined as what « *is available for use in reasoning and for direct conscious control of action and speech* » (Block, 1995). Phenomenal consciousness refers exclusively to the qualitative and subjective aspect of sensory experiences such as seeing, hearing, smelling or having pain. The IIT focuses on the explanation of phenomenology: why is it the case that when I open my eyes it feels something to see? One thing that is almost certain is that it has something to do with the brain. We know this because of observed *correlations* between neural activity and consciousness. Neuroscientists are indeed able to identify the neural correlates of consciousness (Koch, 2004). Modern technology allows to observe with a high level of details which neural networks are involved in the different aspects of conscious experience. However these correlations fail to explain why this neural activity systematically goes together with phenomenology.

In his very influential article *Facing Up the Problem of Consciousness*, David Chalmers argues that cognitive sciences are only able to tackle what he calls the « *easy problems* » of consciousness. The easy problems include the ability to categorize environmental stimuli, the access and reportability of mental states, the focus of attention and the deliberate control of behavior. These are of course very difficult issues that are still subject to active research, but they are easy in the sense that they are « *straightforwardly vulnerable to explanation in terms of computational or neural mechanisms* » (Chalmers, 1995). The *Hard Problem* is specifically the problem of subjective experience. It is hard in

the sense that there exists an *explanatory gap* (Levine, 1983) between physical processes and experience: cognitive sciences do not have the conceptual tools to explain how mechanisms, however complex they may be, can give rise to phenomenal consciousness. Supporters of the IIT claim that it is a fundamental enough theory to bridge the explanatory gap and that it is therefore a valid solution to the Hard Problem of consciousness.

In other words, the IIT is supposed to be addressing the question of subjective experience in a scientific and objective way. More precisely, this implies being able to answer scientifically to questions such as: Why do some brain processes are accompanied by experience and not others? Is it necessary to have a biological brain to be the subject of experience? Do all the other animals have subjective experience? If they do, what are their experience like? Or as Thomas Nagel puts it: « *What is it like to be a bat*? » (Nagel, 1974). Nagel argued that even if we knew everything about the mechanisms of echolocation, we wouldn't know what it's like to experience echolocation in a first-person perspective. Put another way, we wouldn't know the *qualia* associated to echolocation. Qualia refer to the qualitative content of experience accessible by introspection. For instance when I see a red thing, my experience of seeing red has an ineffable qualitative aspect called the quale of red. Similarly, when I am in pain, my experience contains an unpleasant quale of pain. Important expectations regarding a complete scientific theory of consciousness are the necessary and sufficient conditions for a quale to appear and a systematic way to associate objective processes to specific qualia. Therefore, the IIT should be in principle able to derive the qualitative aspects of the subjective experience of a bat from a third-person account of its complete physical constitution.

To accomplish such an ambitious goal, the IIT is based on a specific usage of the notion of information from which is derived the concept of *integrated information*. A formal definition of these notions will be given in the first part of this work. Broadly speaking, integrated information is an abstract quantitative measure of the causal power a system has on itself. The main claim of IIT is the *identity* between informational structures and experience. The nature of this identity will be the subject of the second part. One can interpret the IIT as a fundamental law of nature connecting the physical domain to the mental domain. The philosophical implications of such a claim are numerous and they are subject to criticism. This will be the main concern of the third and last part.

1. What is IIT ?

1.1. A set of phenomenological axioms and ontological postulates

In their article presenting the IIT, the two main defenders of the theory Giulio Tononi and Christoph Koch explain that there are two traditional approaches to consciousness in neuroscience: the Behavioral Correlates of Consciousness (BCC) and the Neural Correlates of Consciousness (NCC) (Tononi and Koch, 2014). BCC is the most common way to assess consciousness for humans, it is based on the inference that if someone behaves as a conscious person, this person is conscious. In particular, if someone is able to report the content of her experience, there is little doubt that she is conscious. NCC is a more sophisticated technique, it consists in observing and interfering with the neural mechanisms associated with conscious experience. For instance by comparing brain imagery during deep sleep and during normal awakened state, a neuroscientist can identify some of the neural activity involved in conscious experience. BCC and NCC are fruitful methods that have greatly increased our knowledge about consciousness, but they are limited. There are indeed a lot of situations where consciousness cannot be assessed with these approaches: dreaming subjects, non-verbal infants, animals or even machines. Moreover, BCC and NCC will never allow to understand *how* and *why* the observed neural patterns give rise to experience. Functional and behavioral explanations can never account for phenomenal consciousness. This has been argued by David Chalmers using the famous zombie argument (Chalmers, 1996, pp.94-96). A zombie is a creature physically identical to a conscious being but lacking conscious experience. It is "nothing like" to be a zombie. According to Chalmers, it is metaphysically possible[1] that I could have a zombie twin: a « *molecule for molecule* » copy of myself, that behaves and talks exactly like I do, but that has no phenomenal experience associated to the material processes happening in his body. The point of the argument is that an explanation of phenomenal consciousness must exclude the metaphysical possibility of zombies.

Starting from the brain and asking how it can produce experience is therefore doomed to failure. In order to hopefully move towards a solution to the Hard Problem, Tononi chooses the opposite direction: the IIT starts with phenomenology and then asks what kinds of physical mechanisms could possibly account for them. The strategy to get

[1] A more detailed account of the zombie argument and its metaphysical implications is provided in section 3.1.

around the Hard Problem is to take experience as *fundamental*[2]*.* Tononi sets out five *phenomenological axioms* concerning the nature of experience from which he derives five corresponding *ontological postulates* concerning the nature of the physical substrate of experience. The five phenomenological axioms of IIT that Tononi considers indubitable are the following (Tononi, 2012 ; Tononi and Koch, 2014):

- Existence axiom: *Consciousness exists*. According to Tononi, this has to be interpreted as a reformulation of Descartes' cogito, replacing the specific "thinking" aspect of experience by the more general notion of having an experience: « *I experience therefore I am* ». My knowledge of the existence of experience is indubitable because I know it immediately from my own *intrinsic perspective*. Consciousness is by definition this intrinsic existence, experienced immediately by any conscious subject. The fact that phenomenal consciousness exists can in fact be questioned and this issue will be discussed in more details in section 2.3.
- Composition axiom: *Consciousness is structured.* This means that experience consists in multiple aspects in various combinations. For example a visual experience contains different phenomenal aspects, like the colors and shapes of objects that *compose* the visual field. Even an experience of pure darkness has a composed content, like the spatial aspect (a left, a center and a right).
- Information axiom: *Consciousness is differentiated,* in the sense that any particular experience exists *by contrast* with other possible experiences. An experience of pure darkness and silence is informative by virtue of the fact that it differs from other possible experiences of colors and sound. This central notion of information will be defined formally in section 1.2. Intuitively, this axiom merely emphasizes the trivial fact that any conscious experience is what it is in part because of what it is *not*.
- Integration axiom: *Consciousness is unified*. The content of experience cannot be reduced to independent components, it is integrated to form a whole. For instance when I experience a visual scene, the object are appearing in a unified field, my visual field. This field cannot be reduced to the sum of separated visual experiences of the objects composing it, perceived objects always appear in a background. An argument in favor of the unity of experience: there is a spatial aspect in a visual field making it possible to separate the left from the right, but by "separating" the right half of the field

[2] The idea that experience is "fundamental" has to be clarified and it will be discussed in more details in the second part of this work.

from its left half, the spatial aspect of the whole experience would be lost. Therefore the whole experience is not a sum of separate "sub-experiences".

- Exclusion axiom: *Consciousness is exclusive.* Experience has definite borders and spatio-temporal grain. I see objects with certain distinctions of colors, and these objects move with a certain temporal grain: an "instant of consciousness" lasts few tens to few hundreds of milliseconds[3]. This axioms states that there cannot be a superposition of another experience with more or less content, so I cannot experience more or less color distinctions at any point in time and I cannot experience the time-flow of consciousness in several superposed way, I never experience "instants of consciousness" of few seconds or few minutes.

IIT posits the five corresponding ontological postulates, concerning the physical substrate of experience:

- Existence postulate: *Mechanisms in a state exist*. This is a minimal ontological claim that merely states that there are such things as "mechanisms" that can be described in terms of evolving states. It is minimal in the sense that for a physical substrate to exist, a necessary condition is that it has causal power[4]. In other words, this postulate is based on the intuition that if experience is real it must be originated by an underlying *causal* process. Denying this would make a theory of consciousness very hard to formulate, at least if one expects such a theory to have explanatory power. An example of a very simple mechanism is a photodiode with two possible states: when it captures enough light it is 'on', otherwise it is 'off'. A neuron is also a binary-state mechanism since it can be described as either 'firing' or 'not firing'.
- Composition postulate: *Mechanisms can be structured.* If experience has a structure, the underlying causal process generating it must also be structured. A mechanism is structured if it has subparts that are themselves mechanisms. For example a photodiode can be part of a larger electronic circuit forming a complex structured mechanism. Another important kind of structured mechanism regarding consciousness is a network of N connected neurons that can be in 2^N possible states (each individual neuron being in one of two possible states).

---

[3] This axiom does not explicitly claim that consciousness is discrete in time. The exact value of time is not important, what is important is that there are no "superposed time-experience" which might sound like a weird and trivial axiom but its relevance will appear more clearly later.

[4] Physical existence and having causal power are taken as synonyms (Tononi, 2017).

- Information postulate: *Mechanisms can generate information.* According to the information axiom, an experience is informative by being different from other possible experiences. In the same way, the corresponding mechanism generates information by being in a particular state, different from other possible states. A neural network of *N* neurons can be in 2^N states. The more states are possible, the more a particular state is informative.
- Integration postulate: *Mechanisms can be unified.* There is a sense in which some structured mechanisms cannot be reduced to the sum of their parts. Let us suppose again a network of connected neurons whose state change over time. Intuitively, if the states of the individual neurons are interdependent in the network, then at any given time I know more about the future state of a neuron if I know the state of the full network than if I only know the state of this single neuron. The extent to which I know more by looking at the whole than by looking only at the parts is quantifiable and this is precisely what *integrated information* refers to. This central notion is presented mathematically in section 1.2.
- Exclusion postulate: *Mechanisms are definite*. A mechanism in a state evolving in time specifies one well defined cause-effect structure. In particular, it has a spatio-temporal grain and it cannot be in a superposition of states.

The axiomatic method on which the IIT is based may pose epistemological problem and is subject to interpretation (McQueen, 2019a). These issues will be discussed in more details in parts 2 and 3.

1.2. A mathematical formalism and a central identity

In order to explain the mathematical content of IIT, the simplest way is to take an example. As explained in the previous section, a mechanism in a state is by definition a causal system that can be composed of connected parts. According to neuroscience, the brain is a mechanism in this sense, since it is made of interacting neurons that can be individually described in terms of states: either the neuron fires (it is 'on', in state *1*) or it is inactive (it is 'off', in state *0*). Let *X* be a mechanism composed of two connected binary-state neurons[5]. *X* can be in four possible states corresponding to a combination of the two individual states of the neuron. The set of possible states of *X* is: {*00, 01, 10, 11*}.

[5] The example that will be presented here comes from (Tsuchiya, 2017).

Let's suppose that the state of *X* evolve in time according to a simple rule: at every time step τ, each of the two neuron copies its state to the other one as illustrated in *Figure 1*.

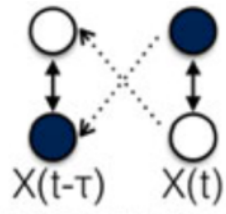

Figure 1. The neural mechanism X represented

in the past ($t - \tau$) and present (t).

Therefore if *X* is in state *01* at time *t-*τ, it will be in state *10* at time *t* and then again in state *01* at time t+τ. If the initial state of *X* is *00* or *11*, the system remains constant. The *entropy*[6] *H* of a system quantifies the uncertainty about the state of a system. If a system *S* can be in *n* possible states $\{s_1, s_2, \ldots, s_n\}$ such that p_i represents the probability that *S* is in state s_i, the entropy *H* of the system *S* is defined as:

(*Entropy*) $$H = -\sum_i log_2(p_i)$$

For example if the state of a system is contant in time, its entropy is 0 because there is no uncertainty about its state. It can be shown[7] that the maximal entropy H_{max} of a system corresponds to an equiprobable distribution, that is when for all i, $p_i = 1/n$, with n the number of possible states.

(*Maximal entropy*) $$H_{max} = log_2(n)$$

X has four possible states, therefore the maximal entropy of *X* is $H_{max}=log_2(4)=2$. Conditional entropy works the same way on the conditional probability distribution of the states of a system. If the evolution of a system is not random, knowing its present state reduces uncertainty about its present and future states. The conditional entropy H^* of *X* is

[6] Entropy is a concept borrowed from Information Theory, a discipline founded by Claude Shannon (Shannon, 1948). However the notion of information in IIT differs in certain ways from the one introduced by Shannon as it will be discussed in section 2.2.

[7] The proof is straightforward, it comes from the concavity of the logarithm function.

the entropy computed on the possible future states of *X* given its present state. In this case $H^*=log_2(1)=0$ because knowing the present state of *X* specifies with certainty one possible state in the future. The mutual information *I* between the present state and the future state is defined as the difference between the maximum entropy and the conditional entropy:

(*Mutual Information*) $I = H_{max} - H^*$

In the case of *X*, $I=2-0=2$. Intuitively, this means that all the information about the future state of *X* is contained in its present state. In the same way, it is possible to compute the entropy of the two neurons of *X* taken separately. In this case there is no mutual information between the present state of a single neuron and its future state. This is because the future state of one neuron depends exclusively on the state of the other one. Let I^* be the sum of the values of mutual information of the subparts of a system. In the case of *X*, $I^*=0$. The *integrated information* ϕ of a system is defined as the difference between the mutual information of the whole system and the sum of the values of mutual information of the parts of the systems.

(Integrated Information) $\phi = I - I^*$

Intuitively, the integrated information generated by a composed mechanism is the quantification of the interconnectivity of its parts. If it is possible to predict with high probability the future states of the parts without knowing the state of the whole system then it means that the parts are not interdependent to a high degree, so the amount of integrated information of the system is low.

A complex mechanism like the brain has a very high associated value of ϕ since it can be in a huge number of states and the state of any individual neuron strongly depends on the states of a lot of other neurons. According to the IIT, this high value of ϕ is precisely what bridges brain activity to conscious experience. One might ask what other type of real-life system has a high value of ϕ. For example, the internet seems to be a network in which the nodes are highly interconnected as its purpose is to exchange messages between any point of the network. However there is a crucial difference between the neural substrate of consciousness and the internet: the latter is not designed to be a

maximum of integrated information. An interaction on the internet is point-to-point, therefore it has to be reducible to independent components by design. If an internet communication involved the whole network it would make the system chaotic. As Tononi puts it: « *the ability to obtain independent, point-to-point signaling excludes the ability to perform global computations, and vice versa* » (Tononi, 2012).

For a better understanding of the difference between the brain and the internet, let's suppose a system of four neurons A, B, C and D. Let's suppose that there is a lot of mutual information between A and B and between B and C but that D is weakly connected to the rest of the network, as illustrated in *Figure 2*.

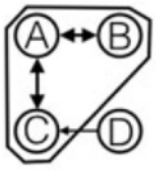

Figure 2. An illustration of the Minimum Information Partition (MIP)
for a simple network of four neurons.

As explained before, it possible to compute the value of ϕ for any part of this network, such as ϕ_{AB} which corresponds to the integrated information of the system composed of the two neurons A and B. For any network, there is one cut called the Minimum Information Partition (MIP) that will minimize the value *I** that corresponds to the sum of the individual values of *I* for each node, as explained before. In the example of *Figure 2*, the MIP cuts the sub-network formed of A, B and C from the neuron D. This means that this MIP is more integrated than the whole network, $\phi_{ABC} > \phi_{ABCD}$. Moreover, there are no sub-network of ABC with a higher value of ϕ. According to the IIT, this implies that in this case the network ABC, and only this network, is associated with a conscious experience. The causal action that the node D exerts on the network ABC corresponds to an unconscious processing. Therefore IIT defends a form of internalism in the context of philosophy of mind: though the brain is causally connected to an external environment, experience ultimately supervenes on the internal state of the brain (or on the internal state of any maximally integrated mechanism).

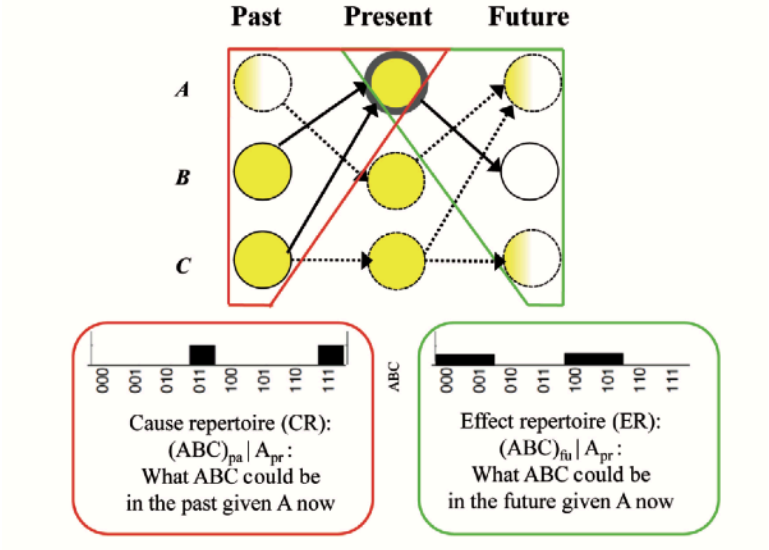

Figure 3. The integrated network ABC defines a cause-effect structure. (Tononi, 2012)

By evolving over time, a maximally integrated mechanism such as the network ABC defines what is called a *conceptual informational structure*. A precise presentation of what is a conceptual informational structure would require a lot of mathematical subtleties that are not necessary for the sake of philosophical discussion about the IIT. To make it reasonably simple to picture, the essential thing to understand is that a time-evolving maximally integrated system defines a mathematical structure in a space with *n+1* dimensions: a time dimension and *n* dimensions corresponding to the *n* possible states of the system. A point in this space therefore defines the exact state of the system at one point in time. This space is called the *qualia space*. The idea is that there is an identity between the content of experience (what is traditionally referred as qualia) and a structure in this space. In other words, the IIT is based on the following claim:

> ***The central identity of IIT:*** *« an experience is a maximally integrated conceptual (information) structure or quale – that is, a maximally irreducible constellation of points in qualia space. »* (Tononi, 2012)

In other words, whenever I subjectively have a conscious experience like hearing a sound, feeling pain or seeing a color, the qualitative aspects of this experience is *identical* to a causal process, the conceptual informational structure. This identity is different from the one claimed by the Mind/Brain Identity theory (Smart, 2019). The identity is not between qualia and the brain, it is rather an identity between qualia and abstract structures

specified by the brain. At first glance, this may look like an intermediate position between the identity theory and functionalism (Levin, 2018). However, the status of this identity will be discussed in more details in sections 2.1 and 2.2 where it will be shown how the IIT is a panpsychist theory of mind.

1.3. A scientific explanation of consciousness

One of the reasons of the popularity of the IIT is its explanatory and predictive power. Since the IIT provides a rigorous mathematical definition of integrated information, it can in principle be computed for any evolving system. Although it is not possible to compute it in practice for real brains because of the high combinatorial complexity of deriving conceptual structures for such a huge mechanism, the formalism accounts for a lot of empirically well documented phenomena in neuroscience:

- The unconscious cerebellum: Smaller than the cerebrum (what is usually called the brain), the cerebellum is an important part of the human neural system. The cerebellum plays an important role in motor function and may also be involved in other cognitive functions such as attention and language. Interestingly, the complete removal of the cerebellum does not affect conscious experience at all, despite the facts that it has four times as many neurons as the cerebrum and that the two are massively interconnected (Lemon & Edgley, 2010). However the cerebellum is composed of modules that process inputs and produce outputs largely independent of each other. Thus the IIT explains why it is not involved in conscious activity: information is less integrated in the cerebrum than in the brain. The latter specifies a maximum of integrated of information and is therefore the only substrate of human consciousness.
- The split brain: According to the IIT, it is possible to create two independent conscious experiences from a single one by disconnecting two parts of its physical substrate. This is exactly what happens in the split-brain experiment (Gazzaniga, 1967). In this famous experiment, patients suffering from epilepsy have the two hemispheres of their brain disconnected after the complete section of the corpus callosum. These patients see an image depicting an object with their right eye such that only their left hemisphere have access to the visual information. After seing the image, when they are asked what they have just saw, they answer that they don’t remember seeing anything. But when they are asked to draw the picture they just saw, they are able to do it correctly. The reason is that language is a cognitive function mainly treated by the right hemisphere that did

not see the image whereas drawing is a task mainly performed by the right hemisphere. According to the IIT, when the corpus collosum is sectioned, there are literally two consciousness flowing in parallel. In principle, if the neurons connecting the two hemispheres are removed progressively, the IIT can predict the exact moment when the value of integrated information for a single hemisphere is greater than the one associated with the whole brain, which will be the moment when consciousness splits into two.

- Sleep: If consciousness is quantifiable, one would intuitively think that the moment in life when it is at its lowest level is during dreamless sleep. Some studies show that, as predicted by the IIT, information is less integrated in the brain during sleep by showing a breakdown of effective neural connectivity (Massimini & al. 2005).
- Brain lesions: The IIT predicts that a brain lesion will make someone unconscious if and only if it severely affects the capacity of the brain to integrate information. Recent studies made using transcranial magnetic stimulation in patients with severe brain damage are consistent with this prediction (Casali & al. 2013).
- Meditation: During meditation, the cerebral cortex is nearly silent. Some theories of consciousness predict that neurons have to be active to contribute to consciousness by signaling represented information (Dehaene & Changeux, 2011). By contrast, the IIT is able to distinct unconsciousness with “naked awareness” corresponding to conscious states without content. Indeed, even if the cerebral cortex is silent, the IIT predicts that it is able to specify a conceptual structure composed of “negative” qualia. As briefly explained in the information axiom of phenomenology in section 1.1, an experience of pure darkness and silence is still an informative experience by being differentiated from other possible conscious experiences.
- Modalities of experience: Experience is organized into modalities such that sight, hearing, smell or touch. There are also submodalities such that color and shape within sight. The IIT predicts that these modalities will correspond to distinguishable subsets of the whole conceptual structure of experience. This is hard to demonstrate in practice, but in principle it is a refutable prediction.

The IIT meets many scientific criteria: it makes precise empirical claims and is therefore refutable in principle. If it is true, its practical applications could be revolutionary. For instance, the theory predicts that we could connect several brains to form a unified super-consciousness. Moreover, research on AI could perhaps use the IIT to address the problem of general artificial intelligence.

2. How should IIT be interpreted ?

2.1. Panpsychism

Panpsychism is the view according to which every physical thing is associated with consciousness. Panpsychism is an old idea of which many variants have been defended in the history of philosophy, notably by Spinoza, Leibniz, Schopenhauer and William James (Mørch, 2019b). A version of panpsychism that has been particularly influential on contemporary philosophy of mind is called *Russellian Monism*. In his 1927 book "*The Analysis of Matter"*, Bertrand Russell wanted to defend a kind of monism by showing that « *the traditional separation between physics and psychology, mind and matter, is not metaphysically defensible* » (Russell, 1927, p.10). Metaphysical monism affirms the indivisible unity of being. In the context of philosophy of mind, it is usually opposed to the dualist doctrine of Descartes, according to which there are two fundamental beings: physical stuff and mental stuff. A popular type of monism is physicalism. According to physicalism, everything is physical and we know the fundamental nature of reality through physics. Russell was not a physicalist in this sense, he argued that physics only tells us about the world as it is *extrinsically*. Indeed, physics allows us to know things through abstract mathematical descriptions, it informs us about the structure and dynamics of object and never tells what these objects are *intrinsically*, in themselves, independently of the effect that they have on us or on our measurement devices. However, as conscious beings, we inhabit the world from an intrinsic perspective. Therefore we know for sure that the extrinsic properties that physics describes are not the only properties that exist. We know it for sure in the sense that it is the most immediate type of knowledge. This argument is based on the Cartesian intuition that experience is an indubitable phenomenon of the world because even doubting about the existence of consciousness requires to be conscious.

Galen Strawson defended a similar argument in his article titled *Why physicalism entails panpsychism* (Strawson, 2006). According to Strawson, physicalism states that every real concrete phenomenon is physical, but it has to be distinguished from physic***s***alism, the view that all we have to know about the physical world comes from physics. Certainly, physicists produce a lot of useful knowledge about matter, but it does not imply that the truth coming from physics are the only truth about the physical world. In particular, it is commonly assumed by physic***s***alists that matter is made of ultimates (such that

elementary particles) that are not in themselves experiential. This assumption is based on the fact that physicists do not need to suppose that the ultimates of the universe are experiential to explain most physical phenomena. However experience is a real concrete phenomenon and physicalism states that every real concrete phenomenon is physical. In addition to everything true physics tells us, we know a crucial property about these physical ultimates that compose our universe: if we put them together in certain ways, there is experience. That raises the problem of explaining experience out of non-experiential beings. This is actually the source of the Hard Problem. One strategy to solve this problem would be to assume that experience "emerges" from non-experiential physical ultimates, the same way as liquidity emerges from non-liquid particles. Yet there is a crucial difference between liquidity and consciousness. The property of being liquid reduces without problem to other properties of physics, it can be easily expressed in terms of shape, size, mass, charge etc. It is not the case of experience, that seems to be *ontologically fundamental*. For example the intrinsic property of feeling pain is something that is not expressible in terms of the extrinsic properties of physics. Therefore, according to Strawson, a real physicalist must abandon the assumption that the ultimates of the universe are in themselves non-experiential. This is why physicalism entails panpsychism. If physicalism is true, the existence of experience at the macro level implies that somehow, experience is a fundamental feature of the universe that already exists at the micro level, in order for emergence to make sense metaphysically.

For anyone taking the Hard Problem seriously, Strawson's argument is quite convincing but his conclusions are counter-intuitive and unclear. Do we have to believe that electrons, quarks and other fundamental particules described by physics are conscious? The fact that it is counter-intuitive is not an argument since a lot of validated scientific theories such that general relativity and quantum physics are very counter-intuitive. Nonetheless, several points about panpsychism need to be clarified. The first question that one might want to ask to a panpsychist is: what does it mean for an elementary particle to be "experiential"? There are different possible answer to that question. Panpsychism does not state that electrons are conscious like humans are, that they have emotions, thoughts or that they perceive the world. The problem that panpsychism is designed to solve is the particular problem of phenomenal consciousness: why are there intrinsic properties, qualia or experiential 'what-it's-likeness' in the universe? The answer of panpsychism is that there must be fundamental entities that are at least "proto-experiential", where proto-experiential means « *intrinsically suited to constituting certain*

sorts of experiential phenomena » (Strawson, 2009). The point is that, whatever they are, the ultimates of the universe must have a particular feature that is not described by physical science, a "micro-phenomenal" property that makes it possible for macro objects like humans to be phenomenally conscious. It could be objected that this statement is meaningless if panpsychists are not able to precise the exact nature of these micro-phenomenal properties. However, it has to be noted that this problem of the nature of phenomenal properties is not specific to micro-phenomenality, it is an issue characteristic to experience in general: we do not know what it's like to be someone else, let it be other humans, other animals or robots if they could be conscious. Therefore it would be probably even more difficult for us to conceive what it would be like to be a fundamental particle. Panpsychism is the metaphysical claim that this property of 'what-it's-likeness' must be a general property of every physical thing[8] even if we will never be able to know the nature of this property.

Another difficulty that panpsychism must face is called the *combination problem* (Seager, 1995). Let's assume that the main claim of panpsychism is true: micro-phenomenal properties exist. How is that a solution to the Hard Problem? How does these micro-phenomenal properties of electrons relate to the qualitative content of my experience? In other words, instead of having to explain how non-experiential material interactions give rise to experience, the panpsychist is facing another version of the Hard Problem consisting of explaining how micro-phenomenal properties combine to give rise to the experiences of colors, pain and hearing. The benefit of panpsychism is that this new version of the Hard Problem is actually not that hard. Of course it is a very complicated issue, but it is not hard in the same sense as in the original Hard Problem because there is no explanatory gap anymore. Indeed, there is no strong emergence (or "brute emergence" as Strawson calls it) of phenomenology from purely non-phenomenal extrinsic properties if everything is experiential or at least proto-experiential. To be complete, panpsychism must account for the general principles or laws governing the micro-phenomenal properties, just as physics accounts for the laws governing the dynamics of elementary particles. For instance, one issue related to the combination problem is whether or not the individual consciousness that aggregate to form a whole

[8] Actually, panpsychism does not make necessary that every physical thing in the universe is experiential but at least that the physical ultimates that conscious creatures are composed of (the fundamental particules that physics describe) are experiential.

conscious experience are lost in the process. Are the individual neurons (or even atoms) that make up my brain conscious ?

This is where the IIT can happen to be useful. According to their defenders:

> *IIT offers a solution to several of the conceptual obstacles that panpsychists never properly resolved, like the problem of aggregates (or combination problem) and can account for its quality.* (Tononi & Koch, 2015)

The IIT can be interpreted as a panpsychist theory of mind. Indeed, according to the IIT, any maximally integrated system is conscious, meaning that a very simple mechanism with only four possible states such that the one illustrated in *Figure 1* (section 1.2) can in principle be conscious since it has a non-zero associated value of ϕ. More precisely, this system will be conscious if it is not part of a greater system with a higher associated value of ϕ. Therefore, the IIT is able to tackle one of the aspect of the combination problem: when a complex mechanism is conscious, its parts are not. This is what happens with the two hemispheres of the brain, as highlighted by the split-brain experiment (explained in more details in part 1.3): each individual hemisphere is intrinsically suited to constituting experiential phenomena but is in general not itself conscious since it is connected to the other hemisphere. The whole brain is conscious but not its subparts. This is a consequence of the Exclusion Axiom of the IIT (part 1.1) stating that two experiences cannot superpose. If a half-brain can be conscious, one might ask if a further division of experience is possible. In theory it is possible, by dividing two parts of one hemisphere that could continue to work independently, although in practice it is probably impossible due to the highly interconnected topology of the brain. The IIT predicts that in principle, the division of the brain-supported experience is possible up to the level of the individual neuron. This is because the description of the brain as a “mechanism in a state” is done at the scale of neurobiology. One could argue that this is arbitrary. A neuron could itself be described as a complex mechanism composed of subparts like molecules. The fact that the IIT only postulates that the physical substrate of experience is a mechanism in a state is both a weakness and a strength. Not knowing the exact nature of what a mechanism is imposes arbitrariness but at the same time it leaves room for interpretation and improvement. In its current form, the IIT is not able to attribute a value of ϕ to an atom or an electron because the

proposed measure of ϕ is only applicable to a network of discrete nodes. However, some physicists already work on the issue of making IIT compatible with fundamental physics, for instance by attributing intrinsic information to fields (Barrett, 2014). Moreover, Tononi and his colleagues treated the special issue of the different "levels of causality" and causal emergence in the context of IIT (Hoel & al., 2013).

2.2. Information, causation and consciousness

This presentation of panpsychism allows to make light on the status of conscious experience in the IIT and its relation to information. According to Russellian Monism, consciousness is associated to all the physical things as it constitutes their intrinsic nature. Another way to put it is to say that physical structures and relations are *realized* by intrinsic properties and that these intrinsic properties have to be identified with experience (Seager, 2006). The underlying assumption is the following: for things to be in relation with each other the way physics is describing them, the relata have to exist in and of themselves independently of the purely relational or dispositional properties that physics reveal. There is an identity between intrinsic existence and consciousness. Hence Russellian Monism avoids both the epistemic gap of physicalism and the causal problem of dualism. Indeed, consciousness is in this perspective neither caused by physical processes nor itself a causal process interfering with the physical causes, it has instead to be interpreted as the realizer of any causal structure. This is an elegant solution as it gives a role to consciousness without breaking the causal closure of the physical world.

However, it can be shown that IIT does not explicitly identify consciousness to an intrinsic property in this sense. The issue comes from an ambiguity in the definition of "intrinsic":

> *An intrinsic property can be defined as a property of a system or entity that does not constitutively depend on properties of other things, or on what is going on in its external surroundings. An extrinsic property can be defined as a non-intrinsic property, or a property of a system or entity that does constitutively depend on properties of other things, or what is going on in its surroundings. An example of an intrinsic property is height because someone's height cannot be changed merely by changing their surroundings. An example of an extrinsic property is "being the tallest person in the room" because this can be changed merely by*

> *changing the person's surroundings, i.e., by adding or removing other tall people to or from the room.* (Mørch, 2019a)

With this definition in mind, Hedda Hassel Mørch argues in her article that by claiming an identity between consciousness and integrated information, the IIT makes consciousness an extrinsic property. She shows this by emphasizing the fact that in the case of the split-brain experiment, the consciousness of one hemisphere depends on the corpus callosum connecting it to the other hemisphere, which is external to both hemispheres. Therefore consciousness is an extrinsic property of the hemisphere because it can be lost and recovered depending on external factors, the same way the tallest person of a room can lose its property if a tallest person enters the room. However, there is a crucial difference between the properties "being the tallest person in the room" and "being conscious according to the IIT" regarding intrinsicality. When the tallest person of a room loses his status because of the arrival of a taller person, there is no causal relation between the two people, whereas the two hemispheres must be causally connected to lose their individual consciousness. The IIT likes causality and intrisicality in a very specific way.

To better understand how the IIT can address this problem of intrinsicality, it is important to first analyze the concept of information that it is based on. Information is a very old concept used in a variety of ways throughout history (Adriaans, 2019). One very influential formalization of information is the one introduced by Claude Shannon (Shannon, 1948). Shannon wanted to solve an engineering problem related to digital communication: given a system that can be in several well-defined states, how to efficiently use a particular state of the system to specify a particular message from the range of all possible messages expressible by that system. He invented the measure of entropy (explained in section 1.2) that is a quantification of uncertainty based on the probabilities of the states of a system. An important aspect of Shannon's information is that it is purely syntaxic: « *semantic aspects of communication are irrelevant to the engineering problem* » (Shannon, 1948, p. 379). Shannon's theory allows to design an efficient digital communication system independently of the meaning of the messages its users will transmit. Therefore this mathematical definition of information is purely extrinsic because nothing can be said to be absolutely or objectively informative in itself. Something is informative in Shannon's sense only relative to a conscious observer that is able to communicate. This is why John Searle argued, in a critique of Koch's book about the IIT, that any theory of consciousness based on information is circular:

> *Consciousness is independent of an observer. I am conscious no matter what anybody thinks. But information is typically relative to observers. These sentences, for example, make sense only relative to our capacity to interpret them. So you can't explain consciousness by saying it consists of information, because information exists only relative to consciousness.* (Searle, 2013)

To answer to these types of objection, the defenders of IIT make it clear that their definition of information differs from Shannon's one:

> *Note that the notion of information in IIT differs substantially from that in communication theory or in common language, but it is faithful to its etymology: information refers to how a system of mechanisms in a state, through its cause–effect power, specifies a form ('informs' a conceptual structure) in the space of possibilities.* (Tononi & Koch, 2015)

In the context of the IIT, being informative is not a relational property between a system and a conscious user, it is an objective feature of a causal structure specified by a system. The causal structure in question is what has been defined in section 1.2 as a "maximally integrated conceptual structure". The IIT is based on a causal definition of information. In this perspective, for something to generate information, it must have causal powers. Moreover, for something to generate integrated information, it must have causal powers *on itself:*

> *From the intrinsic perspective of a system – photodiode or human – information can best be defined as a "difference that makes a difference": the more alternatives (differences) can be distinguished, to the extent they lead to distinguishable consequences (make a difference), the greater the information.* (Tononi, 2012)

This idea of information as a "difference that makes a difference" is taken from Gregory Bateson's causal definition of information:

> *The world of form and communication invokes no things, forces, or impacts but only differences and ideas. (A difference which makes a difference is an idea. It is a "bit," a unit of information.)* (Bateson, 1972, p. 276)

What Bateson means by this definition is simply that something transmits information to another system insofar as it changes the state of this system. This corresponds to the definition of causation, or at least one possible definition of causation. According to the manipulability theory of causation, mainly supported by James Woodward (Woodward, 2003) and Judea Pearl (Pearl, 2000), saying that there is a causal relationship between *X* and *Y* means that we can manipulate *Y* (the effect) by manipulating *X* (the cause):

> *The claim that X causes Y means that, for at least some individuals, there is a possible manipulation of some value of X that they possess, which, given other appropriate conditions (perhaps including manipulations that fix other variables distinct from X at certain values) will change the value of Y or the probability distribution of Y for those individuals.* (Woodward, 2003, p.40)

Therefore, I can consider myself as an information-generating system in virtue of the fact that I have causal power on my environment. Indeed, while I am writing these lines, I am changing the state of my laptop by interacting with it through a keyboard. My laptop also generates information by causally affecting the state of my visual cortex through my retina. Moreover, my own brain generates information by causally affecting itself since its future state strongly depends on its present state. This causal relation that a system has on itself is the definition of integrated information. Therefore it is possible to say that according to the IIT, integrated information and causation are identical, as it is suggested in the most recent presentation of the theory: « *IIT 3.0 explicitly treats integrated information and causation as one and the same thing* » (Tononi & Koch, 2015).

How does all of this make my brain conscious? According to the central identity of IIT, consciousness is integrated information (section 1.2). Since it has just been explained why integrated information can be interpreted as identical to causation, it follows that the central identity of IIT can be re-expressed as follows:

> ***The central identity of IIT****: an experience is a maximally integrated causal structure.*

How can such an identity can make sense? Russellian Monism makes clear that consciousness is the intrinsic nature of the objects described by physics. IIT is not

compatible with Russellian monism because the theory does not postulate explicit ultimates whose intrinsic nature could be identified with experience. An electron cannot be said to be intrinsically conscious since this will depend on whether or not it is part of a grater mechanisms. The IIT postulates the existence of “mechanisms” or “systems” but these notions are themselves defined in terms of spatio-temporal relations of subsystems. For instance when IIT says that my brain is conscious, my consciousness cannot refer to an intrinsic property of my brain since my brain is itself defined as a neural network which is the sum of relations between neurons. Even though IIT makes reference to an “intrinsic perpective of a system”, this intrinsicality is merely epistemic, it cannot be identified with Russell’s metaphysical intrinsic nature of things.

However, there is still a way to make sense of the intrinsic character of consciousness in the IIT by using the concept of phenomenal bonding (Goff, 2016). As a panpsychist, Philip Goff introduced this idea as a way to solve the combination problem. The details of his solution are not relevant for this discussion, what is important is that it is possible to reinterpret an aspect of his idea in the context of the IIT. Goff’s idea was that a physical relation could have an intrinsic nature that is not reducible to the nature of the relata. For example he suggested that the spatial relation could have an intrinsic nature: « *the spatial relation must have some real nature that goes beyond the mathematical conception of it we get from physics* » (Goff, 2016, p. 15). He argued that the spatial relation could be identified with consciousness, involving a universalist variant of panpsychism where every two spatially related things form a conscious entity. In a similar manner, IIT identifies consciousness with a physical relation, that is causality: « *there is an identity between phenomenological properties of experience and informational/causal properties of physical systems* » (Tononi & al., 2014). In other words, consciousness could be interpreted as the intrinsic nature of causality according to the IIT. More precisely, some causal relations have an intrinsic experiential nature. This seems to be the only way to make sense of the claim that IIT’s use of information is both « causal and intrinsic » (Tononi & al., 2014).

This view has interesting implications both in philosophy of mind and in the metaphysics of causation. On the one hand the nature of causation has always been a mystery to philosophers. On the other hand, mental causation is a problem for physicalism, since it is hard to find a place for the mind in a physically closed universe. William James suggested that there could be a link between the nature of causality and the nature of the mind:

> *[…] the concrete perceptual flux, taken just as it comes, offers in our own activity-situations perfectly comprehensible instances of causal agency […] If we took these experiences as the type of what actual causation is, we should have to ascribe to cases of causation outside of our life, to physical cases also, an inwardly experiential nature. In other words, we should have to espouse a so-called "pan-psychic" philosophy.* (James, 1911, p. 218)

This is the argument for panpsychism from experience of causation:

> *I. Non-reductionism: All physical things have causal powers.*
> *II. Mental causation: The only causal powers whose nature we can know, or positively conceive of, are mental.*
> *III. Non-skeptical realism: The nature of the causal powers of physical things is knowable, or positively conceivable.*
> *Therefore,*
> *IV. Panpsychism: All physical things have mental properties.* (Mørch, 2019b)

The second premise is perhaps the most controversial. The idea is that I know causality only from an intrinsic point of view, in the sense that I feel my own causal power. When I decide to lift my arm, I feel that I have causal power on the external world. By contrast, I only infer the causal power of other systems from observation, as Hume famously argued (Hume, 1748). Paradoxically, physicalism claims that I, as a subject, have no causal power but that these external physical things do. The IIT could dissolve that paradox with the following claim: I am myself a causal process. My causal power is therefore not an illusion. It is not an illusory mental causation different from the physical processes. Other causal processes (to the extent that they are maximally integrated) are also conscious. This idea shares common points with the philosophy of Spinoza and that of Schopenhauer:

> *Only from a comparison with what goes on within me when my body performs an action from a motive that moves me, with what is the inner nature of my own changes determined by external grounds or reasons, can I obtain an insight into the way in which those inanimate bodies change under the influence of causes, and thus understand what is their inner nature.* […] *From the law of motivation I*

> *must learn to understand the law of causality in its inner significance. Spinoza (Epist. 62) says that if a stone projected through the air had consciousness, it would imagine it was flying of its own will. I add merely that the stone would be right.* (Schopenhauer, 1859)

However, the panpsychist version of IIT suffers from important flaws that will be discussed in the next section. For those considering the previous explanation as pure metaphysical speculation, there is an alternative way to interpret the theory.

2.3. Illusionism

The first axiom of IIT claims that « *consciousness exists* », without other argument than a reference to Descartes: « *I experience therefore I am* » seems indubitable. But is it really? Keith Frankish recently argued that phenomenal consciousness is an illusion, and he calls this position illusionism:

> *Illusionism makes a very strong claim: it claims that phenomenal consciousness is illusory; experiences do not really have qualitative, 'what-it's-like' properties, whether physical or non-physical.* (Frankish, 2016)

Illusionists claim that there is no Hard Problem because there is no phenomenal properties to explain from non-phenomenal properties. There is of course a sense in which we experience the world, but an "experience" is defined functionally, it has no phenomenal properties such as qualia. An experience is identical to a mental state causally produced by a sensory system, and there is nothing more to it. According to Frankish there are physical properties of experience that introspection misrepresents as phenomenal, that he calls "quasi-phenomenal" properties. For example, I can know by introspection that I am seeing a red book when I am seeing a red book. The nature of introspection deserves to be debated, but the main point is that according to Frankish it is a cognitive ability that can be defined functionally. His idea is that introspection generates the belief that there is something it is like to see a red book, but that this is a wrong belief. Introspection generates the illusion that there is something non-physical to explain.

Instead of the Hard Problem, illusionists need to solve the illusion problem: why do we have the illusion that experiences have phenomenal properties? The main benefit of this view is that a solution to the illusion problem seems much more accessible than a solution to the Hard Problem, with no explanatory gap to be bridged since quasi-phenomenal properties are purely physical. The only thing to explain is how a physical system such that a human being misrepresents his experience as having phenomenal properties. This is an "easy problem" according to Chalmers' classification because it only requires to elucidate the cognitive mechanisms involved in this process of misrepresentation. However, the illusion problem has no straightforward solution because it is not clear how we could have acquired the concept of phenomenology if it corresponds to nothing in reality. Moreover, the illusion of phenomenal consciousness seems to be the most powerful of all the illusions, to the point that most people find the claim of illusionism reluctant or crazy. According to the illusionist François Kammerer, this is actually the hardest aspect of the illusion problem, and the most important to solve in order to make the theory credible and answer the accusations of craziness: « '*It is true that our theory seems crazy, but it explains very simply and naturally many things, including the very fact that it seems crazy.*' » (Kammerer, 2016)

According to Kevin J. McQueen, the IIT could be reinterpreted in order to be a solution to the illusion problem (MacQueen, 2019b). At first glance, it seems that the IIT and illusionism are absolutely incompatible since the explanandum of the IIT is phenomenal consciousness. Moreover, the whole point of the IIT is to start from the indubitable aspects of phenomenology and to rely on introspection to build up its axiomatic foundations. But this does not prevent an illusionist to use the tools provided by Tononi to explain why humans have the illusion the have a phenomenology without having to deal with the very counter-intuitive implications of IIT about what has to be called "conscious" since an illusionist thinks that nothing in the universe is conscious. In order to do this, McQueen proposed to reformulate the axioms of IIT to define what he calls Illusionist-IIT:

> ***Illusion Axiom****: experiences have quasi-phenomenal properties, where a quasi-phenomenal property is a non-phenomenal property of experience that introspection typically misrepresents as phenomenal.*
>
> ***Intrinsic Existence Axiom****: quasi-phenomenal properties are introspectively represented as being intrinsic properties that are private and immediately known.*

Composition Axiom*: quasi-phenomenal properties are introspectively represented as being composed of many phenomenological distinctions.*

Information Axiom*: quasi-phenomenal properties are introspectively represented in terms of informative phenomenological differences. That is, each experience appears to be the particular way it is by differentiating itself from what it is not.*

Integration Axiom*: quasi-phenomenal properties are introspectively represented as being unified in the sense that each experience is irreducible to non-interdependent subsets of phenomenological distinctions.*

Exclusion Axiom*: quasi-phenomenal properties are introspectively represented as being definite in content and spatio-temporal grain.* (MacQueen, 2019b)

Moreover, the five corresponding postulates also have to be redefined:

Illusion Postulate*: to support quasi-phenomenal properties, a system must have the type of physical property that is systematically introspectively misrepresented as being phenomenal.*

Intrinsic Existence Postulate*: to support quasi-phenomenal properties, the system must have intrinsic causal power. This means having causal power over itself, independently of external factors.*

Composition Postulate*: to support quasi-phenomenal properties, the system must be structured into parts that themselves have causal power within the system.*

Information Postulate*: to support quasi-phenomenal properties, the system must specify a causal structure that differentiates its state at one time from its state at other times. That is, it must contain information about itself.*

Integration Postulate*: to support quasi-phenomenal properties, the system must be unified, or irreducible to a simple sum of component causal structures. That is, the self-information it contains must be integrated.*

Exclusion Postulate*: to support quasi-phenomenal properties, the system's causal structure must be specified over a single set of elements, the set that yields the maximum amount of integrated information.* (MacQueen, 2019b)

There are several reasons why Illusionist-IIT might be more appealing than the orthodox version:

- Illusionist-IIT does not require complicated investigations such that the one carried out in the previous part about the status of phenomenal properties, their intrinsic character and the link between information, causation and experience. It is indeed difficult to make sense of the metaphysical claim that qualia are “intrinsic” properties. By contrast, quasi-phenomenal properties are not intrinsic, they are only represented as intrinsic.
- IIT is based on a controversial epistemic claim: phenomenal properties are private and known with immediate certainty through introspection and it can be argued that introspection if often misleading. This can be considered as a major flaw and is avoided by Illusionist-IIT.
- IIT specifies a set of axioms that most would consider necessary to describe the nature of consciousness. However, this set could be incomplete, nothing says that they are sufficient conditions to qualify an experience. It may be precisely because this set is incomplete that the IIT attributes experiential properties to systems that are intuitively not conscious. Illusionist-IIT merely needs necessary conditions to explain why advanced cognitive systems misrepresents their experience, with a concept of experience defined functionally and only attributable to very specific creatures (humans and maybe other animals).
- Illusionist-IIT is based on the very counter-intuitive claim that phenomenal properties do not exist but in counterpart it avoids IIT’s most counterintuitive prediction that some very simple electronic system are conscious. Another major flaw of the IIT concerns the theory’s prediction that a 2D grid of around one billion squared identical logical XOR gates can have the same value of ϕ than a humain brain. According to objectors to the theory, no acceptable definition of quantified consciousness can be such that a system with low computational complexity is attributed as much consciousness as a human being (Aaronson, 2016a, 2016b). Illusionist-IIT attributes consciousness to nothing, a grid of logic gates is in this perspective trivially unconscious. Moreover an electronic system is not capable of introspection, therefore it does not have the illusion of having a phenomenology.

Thus, Illusionist-IIT is able « *to solve several problems that plague IIT* » (MacQueen, 2019b). How can it solve the problems of illusionism? A complete answer to this question would go beyond the scope of this work but some relevant points can be noted:

- As explained before, one aspect of the illusion problem is to find out where our defective phenomenal concepts come from. When Einstein proposed his theory of relativity, physicists discovered that their concept of aether referred to nothing in reality. However they knew why they believed it existed, aether was supposed to be a material substance in which light could travel, and it made sense to suppose it. Similarly, illusionists want to find the physical origin of the phenomenal concepts, an approach sometimes called the phenomenal concept strategy (PCS). As David Chalmers put it, PCS « *locates the gap in the relationship between our concepts of physical processes and our concepts of consciousness, rather than in the relationship between physical processes and consciousness themselves* » (Chalmers, 2006). The axioms of IIT could be interpreted as specifying some aspects of our conception of phenomenology, such that being integrated, informative and unified. The corresponding postulates could thus make the link between the original concepts referring to physical systems (such that being integrated) and the defective phenomenal concepts. In other words, some systems such that human beings think that their experience have a rich content with intrinsic properties because they are themselves integrated systems.
- It could be asked why these concepts are so useful if they are defective. A possible answer could be that generating integrated information could be evolutionary advantageous (Popiel & al., 2020). This explanation is not specific to Illusionist-IIT, it is an argument for IIT in general. In the specific context of Illusionist-IIT, one explanation could be that it was useful for us to evolve internal representations of experiences to communicate them. What these representations are truly representing are very complex brain states, therefore they had to be introspectively accessible in a distorted way that made us think that they have special properties. The general idea would be that we had to compress the information in a useful way that made communication possible. Of course, this is very speculative and deserves a more developed argumentation based on empirical facts.
- Finally, the hardest aspect of the illusion problem, namely the powerfulness of the illusion, could be related to the value of ϕ. IIT could be reinterpreted as follows: « *the strength of a subject's introspective illusion is proportional to the* Φ^{max} *of the introspected state* » (MacQueen, 2019b).

3. Can IIT be a solution to the Hard Problem of consciousness?

3.1. IIT as a scientific theory of phenomenal consciousness

In his 1996 book, Chalmers presented the zombie argument as « *the most obvious way (although not the only way) to investigate the logical supervenience of consciousness* » (Chalmers, 1996, p.96). One of the main claim of physicalism is that consciousness supervenes on physical states, which means that for any conscious state to possibly change, a physical change is necessary. Chalmers argued that there are two kinds of supervenience (Chalmers, 1996, pp. 34-38):

Logical supervenience: « *B-properties supervene logically on A-properties if no two logically possible situations are identical with respect to their A-properties but distinct with respect to their B-properties.* » For example, biological properties logically supervenes on physical properties because fixing all the physical facts about the universe logically entails that all the biological facts are fixed. « *Even God could not have created a world that was physically identical to ours but biologically distinct*. »

Natural supervenience: This corresponds to a supervenience regarding the laws of nature. Chalmers claims that we could imagine a world with different laws of nature, so some properties like the pressure of a gas supervenes on its temperature but it could be, with different laws of nature (different constants for instance) that it is not the case.

Chalmers' point is that consciousness does not logically supervenes on the physical world. Given all the physical facts, one could not deduce that consciousness exists. Put in a more metaphysical way, God could have created an identical physical world with no consciousness. This identical world is called the zombie world. In the zombie world, everything is molecule for molecule identical to our universe. For example, in the zombie world I have a zombie twin that is physically and behaviorally identical to me. He does and says the same things that I do. The only difference is that it is nothing like to be my zombie twin, it is « dark inside ». When my zombie twin undergoes a physical damage, he behaves in such a way that he is in pain, but he does not subjectively feel pain because in the zombie world there are no qualia. The zombie world is conceivable because according to physicalism, my behavior is functionally determined by a physical causal substrate. Everything could happen exactly the way it happens with no qualia involved.

Therefore the question is: why isn't our world the zombie world? In Chalmers' mind, a theory of consciousness that is designed to solve the Hard Problem should be able to answer this question.

Does the IIT explain why we are not living in the zombie world ? According to the IIT, consciousness is identical to integrated information and the latter is itself defined as identical to causal structures generated by physical mechanisms. Therefore, saying that consciousness is integrated information does not seem at first glance very different from saying that consciousness supervenes on the physical world. It is conceivable that integrated information is not identical to consciousness. I could still have a zombie twin that have the same value of ϕ attributed to his brain. Therefore, it could be concluded that the IIT does not solve the Hard Problem of consciousness because the zombie argument works against the IIT. However, understood in this sense, the Hard Problem is intractable by design. Whatever the theory bridging physical processes to consciousness, one could conceive that the theory is false and conclude that it is not a valid explanation. This is why the distinction between logical and natural supervenience is important. The identity between consciousness and integrated information is not a logical identity, it is a natural identity: it happens that in our world some causal structures are qualia. Conscious states naturally supervenes on the causal structures specified by integrated mechanisms the same way the pressure of a gas supervenes on its temperature.

Another way to put it is to use Saul Kripke's distinction between logical and metaphysical possibility. In *Naming and Necessity*, Kripke argued that it is logically possible that water is not H_2O (Kripke, 1980). We could imagine a world where "watery" stuff is identical to another molecule, like XYZ. But we discovered that in our world, water is H_2O. Therefore, according to Kripke, the identity between water and H_2O is an a posteriori necessary truth. Its negation is logically possible but not metaphysically possible. In that respect, IIT's central identity between conscious states and maximally integrated causal structures can be interpreted as an a posteriori necessary truth. It is certainly logically possible that IIT's identity is false, but it is not metaphysically possible. In a sense, Tononi discovered that consciousness is integrated information, the same way that physicists discovered that water is H_2O, that PV = nRT or that gravity is a curvature of spacetime. This would give IIT's central identity the status of a fundamental law of nature.

Could Tononi's IIT be a fundamental theory of consciousness in the same sense as Einstein's theory of general relativity is a fundamental theory of gravity? The two theories certainly have common features. They both are mathematically precise formalisms and they both defy common sense. Indeed, Einstein's theory of relativity claims counter-intuitively that we live in a curved four-dimensional structure and that the flow of time is relative to the velocity of an observer. Moreover, general relativity affirms that gravity is not to be interpreted as a force exerted by material bodies as Newton thought, but rather as a complex geometrical transformation. Analogously, IIT claims that consciousness is not a biological function of nervous systems as we thought, but has instead to be interpreted as a geometrical structure in an abstract qualia space. In both cases, there is a sense in which the explanation is incomplete: if one is tempted to ask why it is the case that consciousness is integrated information, she would analogously have to ask why material objects curve spacetime. Both theories are claiming to provide a posteriori necessary truths about our universe, and further metaphysical questions such that "why is the universe the way it is?" are not in their scope.

Nevertheless there is an important difference between a theory of gravity and a theory of consciousness: we know what gravity refers to but it is not clear at all in the case of phenomenal consciousness. The main difference is that gravity is an objective phenomenon, in the sense that it is completely revealed by a third-person account: everybody can observe the effect of gravity when objects fall down or when the sun moves during the day. A scientific theory can be broadly defined as an explanation of an objective observable phenomenon. The problem with a scientific theory of phenomenal consciousness is that phenomenal consciousness is not an objective observable phenomenon, it is irreducibly subjective and is not observable because it is itself the condition of possibility of any observation. As John Searle puts it:

> *We find it difficult to come to terms with subjectivity, not just because we have been brought up in an ideology that says that ultimately reality must be completely objective, but because our idea of an objectively observable reality presupposes the notion of observation that is itself ineliminably subjective, and that cannot itself be made the object of observation in a way that objectively existing objects and states of affairs in the world can. There is, in short, no way for us to picture subjectivity as part of our world view because, so to speak, the subjectivity in question is the picturing.* (Searle, 1992, pp. 97-98)

The most problematic consequence of this irreducibility of the subjective is that there is no way to measure consciousness. The IIT proposes a way to quantify consciousness with the measure of ϕ but the whole operation is circular because the fact that ϕ is a correct measure of consciousness is based on the presupposition that consciousness is integrated information. IIT predicts that any system with an associated value of ϕ is "conscious" but there are no objective facts that could in general confirm or infirm the existence of intrinsic phenomenal properties of a system. One particularly problematic example is the case of the conscious 2D grid of identical XOR gates, as already discussed in part 2.3. IIT predicts that with a sufficient number of logical gates[9], the quantity of integrated information generated by the system could be the same as that of a human, suggesting that such a system is as "conscious" as a human being, a very counter-intuitive prediction that Tononi coherently accepts (Tononi, 2016). For most people, such a prediction only indicates that IIT's definition of "consciousness" is defective[10]. The source of the problem is that in order to know what is a phenomenal conscious state, one has to *be* in that state. This was argued by Frack Jackson in his famous article *What Mary Didn't Know* (Jackson, 1986). A scientist that has never subjectively experienced red does not know what is the phenomenal property of "red", even if she knows every physical fact about the visual experience of red, such that surface reflectance properties and neural network mechanisms happening in the visual cortex. By contrast, she would know everything relevant about gravity by reading a book about general relativity. Of course, one could argue that if Mary never experienced gravity, for instance if she always were in an anti-gravity room without windows, she wouldn't actually know everything that has to be known about gravity. But in this case what she wouldn't know would be "what it's like" to feel gravity, which would again refer in a way or another to the concept of phenomenal consciousness.

Therefore there is a sense in which a complete scientific account of consciousness is impossible, even in principle. In his book *Galileo's Error*, Philip Goff argues that « *the*

---

[9] It would require billions of billions of logical gates which in practice would be very difficult to wire.

[10] For instance, the computer scientist Scott Aaronson does not believe at all that IIT is a serious candidate for a theory of consciousness: « *On reflection, I firmly believe that a two-state solution is possible, in which we simply adopt different words for the different things that we mean by "consciousness"—like, say, consciousness$_{Real}$ for my kind and consciousness$_{WTF}$ for the IIT kind.* » (Aaronson, 2016b)

scientific revolution itself was premised on putting consciousness outside of the domain of scientific inquiry » (Goff, 2018, p.37). Often considered as the father of modern science with his revolutionary idea that « *the grand book of nature is written in the language of mathematics* », Galileo would have found absurd the project of finding a scientific explanation of consciousness. Galileo's project was revolutionary precisely because he withdrew the sensory qualities such that color, smell, taste and sound from the purely quantitative domain of science:

> *Galileo's universe was divided up into two radically different kinds of entity. On the one hand, there are material objects, which have only the mathematical characteristics of size, shape, location, and motion. On the other hand, there are souls enjoying a rich variety of forms of sensory consciousness in response to the world. And the benefit of this picture of the world was that the material world with its minimal characteristics could be entirely captured in the language of mathematics. This was the birth of mathematical physics*. (Goff, 2018, p.39)

The roots of the Hard Problem have to be found at the origin of modern science. The bottom line is that there seems to be a categorical difference between the qualitative aspect of the content of experience and the quantitative account that any fundamental scientific theory is able to provide.

3.2. IIT as a bridge between the qualitative and the quantitative

In addition to the irreducible subjective ontology of consciousness, a complete solution to the Hard Problem must tackle the qualitative aspect of phenomenal consciousness. What is exactly this so-called "qualitative" aspect that makes qualia so difficult to explain? As explained in the previous section, Galileo opposed the domain of the qualitative to that of the quantitative. In this sense, what counts as a qualitative property is defined negatively as what is not captured by mathematics. However one could object that this is not a satisfactory definition of qualitative as most people would consider shape as a physical qualitative property of objects that is fully captured by geometry. To better understand in what sense there are some properties that are not captured by mathematical approaches, philosophers often use the paradigmatic example of colors. Unlike perceived shapes, perceived colors are thinkable independently of colored objects. The best way to show this is through the conceivability of inverted spectrum (Shoemaker, 1982). The visible

spectrum associates perceived colors to wavelength, for instance the lowest wavelengths (≃ 400 nm) are associated to violet and the highest ones (≃ 700nm) are associated to red. Violet and red can refer to objective colors that are quantifiable physical properties. The fact that a human visual system is able to distinguish the two colors can be explained functionally. By contrast, the color qualia “violet” and “red” refer to “what it’s like” to see violet and red. I can conceive that I could see “violet” what I call red and vice-versa, which emphasizes that there is some arbitrariness involved in the way perceived colors are associated to wavelength: the functional role of distinguishing colors would be preserved even if my perceived spectrum were inverted.

The arbitrariness of perception is a broader question: why is it like it is to hear, smell, taste and feel pain? One can have the intuition that after cognitive sciences provide all the functional explanations, there remains a substantial question about the nature of perceptual experiences. This additional substantial issue is by definition the “qualitative” aspect of phenomenal consciousness and a theory of consciousness have to provide the tools to tackle it. According to some of its defenders, the IIT allows the start of a solution to this difficult problem thanks to its mathematical framework. In fact, the main purpose of IIT is to bridge the gap between the qualitative aspect of experience and the functional role of its physical substrate. Indeed, IIT starts with phenomenological axioms describing the main aspects of experience and then looks at what could account for them in the physical world. Being conscious “feels like” being in a composed and informative unified state, so Tononi infers that the physical substrate of consciousness is also composed, informative and unified. Understood this way, the identity between an experience and a maximally integrated structure is explanatory rather than metaphysical. The aim of the theory is to bridge the explanatory gap by finding a one-to-one correspondance between phenomenal properties and physical properties. Unlike the zombie problem, the inverted spectrum issue does not require to postulate any metaphysical identity, it can be solved by finding the relevant bijection associating color qualia to informational structures.

An interesting approach to translate the qualitative aspects of phenomenology in terms of informational structures is to use category theory (Awodey, 2010 ; Marquis, 2019). Category theory, introduced in 1945 by Samuel Eilenberg and Saunders Mac Lane, is the branch of mathematics concerned with structures and relations at the most abstract level. It is now considered as the foundation of mathematics, replacing set theory that used to hold this position. Category theory formalizes the common definition of a category which

refers to a class of things that share certain properties, such as the category of humans or the category of trees. Combined with IIT, it has been proposed that category theory could be a useful tool to approach the mind-body problem:

> *This formalization of categories allows us to compare and relate two seemingly complete separate domains of knowledge, such as mathematical concepts and conscious experience.* (Tsuchiya & al., 2015)

Category theory provides an abstract framework in which the broad notion of "analogy" is formally definable. For example, it can be easier to solve a geometry problem by translating it into the domain of algebra. It is the case for the Brouwer's fixed-point theorem stating that any continuous point-to-point mapping from a disk to itself will leave at least a single point unchanged (for instance a rotation will leave the center of the disk unchanged). This is a very difficult to prove by using only the concepts of disks and geometrical transformations. However, category theory can show that this problem is identical to a much easier algebraic problem dealing only with transformations on the set of integers. The identity is merely epistemological, not metaphysical. Category theory does not state that there is an identity between the categories themselves but between the structures of the categories. A disk is certainly not identical to the set of integers but it shares certains structural properties that can be useful for explanatory purposes. Recent works have shown that category theory is able to reveal analogies between disciplines such that quantum physics and topology (Baez & Stay, 2009). Category theory offers two important conceptual tools to bridge different domains:

> *First, category theory provides a mathematical framework for translating a relationship in one domain to a distinct and separate domain by use of a structure-preserving map, or a functor. Second, category theory brings a precise mathematical formalism to assess whether or not two separate domains of knowledge are similar and in what qualitative way they are similar.* (Tsuchiya & al., 2015)

This is why category theory and the IIT could be useful in order to precisely analyze the qualitative character of consciousness:

> *We believe that finding a functor between IIT's mathematical structure and consciousness might be also sufficient to bring about many theoretical and empirical results, without requiring "identity" as claimed by the original theory. Before the invention of category theory, there was no systematic framework to characterize this kind of qualitatively graded levels of similarity. From a category-theoretic point of view, strength of similarity, analogy, metaphor, and relationship that are used in many different scientific disciplines can be qualitatively characterized in a very precise manner.* (Tsuchiya & al., 2015)

Therefore, contrary to what Galileo would have expected[11], there may be a way to study sensory qualities in a mathematical way. This is just because mathematics is not purely "quantitative": geometry, topology and algebra deal with forms, structures and relations that are "qualitative" in some sense. According to Tononi, the qualitative aspects of the different modalities of perception have counterparts in the qualia space (Q):

> *We recognize intuitively that the way we perceive taste, smell, and maybe color, is organized phenomenologically in a "categorical" manner, quite different from, say, the "topographical" manner in which we perceive space in vision, audition, or touch. According to the IIT, these hard to articulate phenomenological differences correspond to different basic sub-shapes in Q, such as grid-like structures and pyramid-like structures. In turn, these emerge naturally from the underlying neuroanatomy and neuronal activity patterns.* (Balduzzi & Tononi, 2009)

With a better understanding of the geometry and topology of the information structure specified by integrated mechanisms, partisans of the IIT hope to elucidate Nagel's famous question "what is it like to be a bat?":

> *Although it may be practically impossible to understand bats' phenomenology in every detail, the research project I outlined above would be sufficient to give a highly credible answer as to whether the bat's echolocation is closer to audition, vision or nonconscious processing. Identifying the neuronal connectivity in bats' brain and understanding their neural activation patterns, analysed according to the*

[11] At least, according to Philip Goff's interpretation.

> *IIT's principles will give a fairly educated and grounded answer, assuming that IIT is correct.* (Tsuchiya, 2017)

Independently of the mathematical details, these considerations are important in at least one respect: a scientific approach to consciousness does not necessarily have to suffer from the metaphysical burden outlined in section 3.1. The view that a theory of consciousness must have a metaphysical status comes from the formulation of the Hard Problem and from a metaphysical interpretation of the laws of nature. In section 3.1, it was shown how the zombie argument was forcing to look for a fundamental law that could exclude the possibility of zombie. Physics is often considered as the scientific discipline whose aim is to discover "fundamental" and "universal" laws of the universe such as the law of gravity. However, this is not the only way to view natural sciences. One could adopt an empiricist point of view according to which the purpose of science is to find regularities and symmetries in the natural world, and not "laws" in a metaphysical sense. This is a view notably defended by Bas van Fraassen who proposed a deflationist view of laws of nature (van Fraassen, 1989, 1991). Van Fraassen argued that what we call laws of nature are in fact symmetries, a notion that he formally defines using abstract algebra. He sums up his view by the following simple idea:

> *Symmetry Requirement: Problems which are essentially the same must receive essentially the same solution.* (van Fraassen, 1989, p.236)

Van Fraassen's notion of symmetry is very close to the concept of functor defined by category theory. In both cases these notions can be intuitively understood as isomorphisms between domains. In this perspective, laws are not to be interpreted as transcendant universals concerning an independent reality, they are more modestly model-dependent truths. The framework inside which they are true have to be judged on the basis of their explanatory and predictive power. This pragmatic view of science allows to take IIT's central identity as an abuse of language: just as gravity is not really "in itself" a geometrical transformation of spacetime, an experience is not really identical to an informational structure, but the latter might be the best way to describe it.

However, one might object: if IIT is merely a "bridge" between the qualitative domain of consciousness and the mathematical domain of structures and mechanisms, doesn't this

suggest that there are really two fundamentally different things to bridge? Isn't this view suggesting a form of dualism?

3.3. The ontological Hard Problem

In order to solve the Hard Problem, a theory of consciousness has to be ontologically clear. What really exists according to IIT? Tononi wrote an article specifically about « ontological considerations » (Tononi, 2017). Interestingly, one of the part of the article is titled « Different kinds of existence », where Tononi lists different kinds of things that exist in different ways: intrinsic entities (experiencing subjects), extrinsic entities (physical objects), aggregates (non-conscious "objects" made of smaller objects, like a heap of sand), arbitrary collections (like the set of all beer cans in North America), imaginary objects (like a phoenix) or abstract objects (like numbers). For each type of object, Tononi explains how IIT accounts for them. But this list of existents is based on an ordinary and naïve definition of existence:

> *However, as science has revealed many times, our naïve catalog of entities is often inaccurate and sometimes wrong. For example, we naively think that water is a homogeneous substance, while modern physics shows that its structure is much more complicated, being made of molecules, and these of atoms, which themselves are made of a wide variety of particles and fields.* (Tononi, 2017)

In other words, Tononi argues that it is the role of science to tell what really exists and what exists only as a matter of speaking. But in the case of consciousness, this view is problematic since we do not have a consensual theory of consciousness. The whole point of a scientific theory of consciousness is to reveal the nature of consciousness: what is it really? Is it this intrinsic existence that we intuitively think it is or is it really something else? IIT's answer is simple: consciousness exists, there is no possible doubt about it because we know it immediately.

As discussed in part 2, IIT has to be combined with panpsychism to be metaphysically straight and solve the Hard Problem. The theory leads to postulate the existence of micro-experientiality and identify the "ultimates" of experience with the intrinsic nature of some causal relations. An illusionist would be right to object that it looks a lot like a convoluted ad hoc explanation. It would arguably be much simpler to just say that what

“really” exist are the informational structures and that the theory only shows that intrinsic unified existence is just an illusion coming from the high level of integration of some particular mechanisms emerging from natural selection. In this case the answer to the Hard Problem would be simple: we are actually zombies and it is perfectly coherent because if zombies are functionally identical to us they would ask themselves the same questions about their supposed phenomenology. Phenomenology is an illusion, there is nothing over and above the functional.

The idea that phenomenology is an illusion is generally considered unacceptable. Indeed, the mere existence of an illusion seems to presuppose a form of intrinsic existence. If I do not exist as a conscious subject, how could I even have an illusion? Illusionism has to answer very difficult questions about the exact nature of this phenomenal illusion, as discussed in part 2.3. Nevertheless, this shows that even assuming that the IIT is a “true” theory of consciousness in some sense, the Hard Problem would not be definitely solved because the theory would still have to be interpreted by philosophers. On the one hand, some would claim that the IIT shows that consciousness is almost everywhere, and on the other hand objectors would argue that the theory shows that consciousness is actually nowhere. In the same way as there are several competing interpretations of the quantum formalism that claims completely different things about the nature of the physical world, even an empirically correct theory of consciousness would lead to disagreements about the nature of experience and its ontological status.

In this case, would there be objective facts that could lead one to decide between Orthodox-IIT and Illusionist-IIT? In other words, do such ontological questions have definite answers with determinate truth values? Chalmers himself argued that they probably do not. In an article titled Ontological Anti-Realism, he explains:

> *The basic question of ontology is ‘What exists?’ The basic question of metaontology is: are there objective answers to the basic question of ontology? Here ontological realists say yes, and ontological anti-realists say no.* (Chalmers, 2009, p.77)

Chalmers shows that the disagreement between ontological realists and ontological anti-realists is about the possibility of an “absolute existential quantifier”. Either there is a true general definition of “existence” or there are just different ways of speaking in which case

ontological debates are just verbal disputes. Chalmers takes the example of a mereological disagreement: if two objects exist, do the sum of the two objects also exists? The universalist will say 'yes' and the nihilist will say 'no'. The consequence is that a nihilist will say that "tables exist" is false, he would argue that only the fundamental particles that the table is made of can be said to "exist" in a strong ontological sense. If there were a non-defective absolute quantifier, there would be a way to assign a determinate truth-value to the proposition "if x and y exist, the sum of x and y also exist". But this proposition is not analytic, it cannot be true merely in virtue of the concepts involved: "existence" is a primitive concept that cannot be analyzed and there is nothing in the concept of sum that could elucidate anything about existence. If it is not an analytic it means that no a priori reasoning can refute nihilism or universalism, which means that both nihilism and universalism are conceivable and therefore possible according to Chalmers[12]. But universalism and nihilism are not both possible, since they are logically incompatible. Therefore, there cannot be an absolute existential quantifier.

Chalmers' crucial point is that existential propositions are only true inside a correctly defined domain that is associated to a world[13]:

> *Intuitively, a domain is a catalog of entities that are taken to exist in a given world. To a first approximation, we might model a domain as a class of singular terms in an idealized language.* (Chalmers, 2009, p.107)

The conclusion of the article is that there are good arguments in favor of ontological anti-realism. There are some existential propositions that are true in a domain and false in another, while both domains are admissible. The criteria for what is considered an admissible domain could be further discussed. One requirement would be that a domain must be internally coherent (no existing round square for example). The general idea is that there are cases where there is no objective facts that can favor one domain over another. The criteria to decide which domain is the best are contextual, it usually depends on a lot of practical factors that are not directly related to truth in a strong metaphysical sense.

---

[12] The "conceivability argument".

[13] This is very close to Rudolf Carnap's notion of "framework" inside which truth-values can be assigned. Carnap distinguishes internal and external questions, the latter being empty of cognitive content. (Carnap, 1950)

Ontological anti-realism is relevant in the context of philosophy of mind. There may be no substantial truth to know about the existence of consciousness. It is not to say that the debate between panpsychists and illusionists is only a verbal dispute, but rather to point out that one source of disagreement might be a matter of definition of both “existence” and “consciousness”:

- Panpsychists have such a liberal definition of “consciousness” that it loses sight of the original “what it’s like” problem that may be a very human (or animal) feature. Moreover their definition of “existence” lacks clarity because they want the benefits of both monism and dualism by supposing one fundamental existence with two “aspects”: extrinsic existence is synonymous to causal power and intrinsic existence allows to find a place for consciousness. This is certainly elegant but the notion of intrinsic existence seems to rely on nothing more than an intuition (that could be an illusory human intuition).
- Illusionists, by affirming that we are zombies, lose sight of an important aspect of the Hard Problem which is that it is like something to be awake whereas it is “dark inside” during dreamless sleep. It is hard to articulate a coherent view based on the claim that in fact it is really “dark inside” when we are awake but that we have the illusion that it is not. A definition of “existence” that excludes what makes possible to even think about existence is arguably a problematic definition.

Conclusion

The IIT is a very promising theory of consciousness in many respects:

- It is the first scientific theory of consciousness that explicitly addresses the Hard Problem. Even if the theory is wrong, it is interesting insofar as it gives a glimpse of the limits of such an ambitious project.
- Although some aspects of the theory are subject to interpretation, it is very precise. Both the epistemology and the ontology of IIT are clearly defined. The mathematical apparatus is also well-defined and allows to make precise and refutable predictions which means that in principle some experiments could show that IIT is wrong. If the theory turns out to be false, it will have been precisely wrong, which is an important feature of scientific theories.

- Integrated information is certainly a fundamental neural correlate of consciousness. Maybe it is nothing more than a correlate, a necessary but not sufficient condition for conscious experience. But at worst, the theory provides useful tools to analyze mathematically what it takes for a structure to be integrated. This could have a lot of practical applications in the fields of artificial intelligence and neuroscience.

The major limitation of any theory of consciousness is that what it studies is not a phenomenon in the usual sense, it is the condition of possibility of any phenomenon. IIT's axiomatic method is questionable but it may be the only way to move forward on this very particular problem: taking the existence of phenomenal consciousness as a starting point and carefully describing what it's like to be conscious, then looking at what can account for this in the physical world. In a sense, one that takes the Hard Problem seriously has to endorse a form of dualism, at least methodologically. Indeed, the Hard Problem considers as a premise that there are two very different kind of things that have to be "bridged". As it has been explained, the main reason has to do with causality: physical existence means having causal power whereas qualia are defined as what cannot be reduced to the causal interactions.

Thus the Hard Problem is essentially a metaphysical problem. It is not possible to limit oneself to an empirical point of view to solve it. Panpsychism is an attractive solution to the Hard Problem, and combined with IIT it allows new ways of thinking about the universe that have to be further developed. Panpsychists such as Russell and Strawson are right to emphasize that we know very few about the nature of the physical world. Why then would we want to attribute intrinsic existence to things and relations we do not clearly understand? A complete scientific theory of consciousness capable of solving the Hard Problem would probably require a much better understanding of physics, neuroscience, and perhaps mathematics and computer science. IIT should therefore be considered as a first step towards a fundamental theory of consciousness. What it shows it that philosophy is needed to solve this Hard Problem perhaps more than anywhere else in science.

Bibliography

Aaronson, S. (2016a). “Why I Am Not an Integrated Information Theorist (or, The Unconscious Expander).” The Blog of Scott Aaronson. Available online: https://www.scottaaronson.com/blog/?p=1799

Aaronson, S. (2016b). “Giulio Tononi and Me: A Phi-nal Exchange.” Shtetl-Optimized: The Blog of Scott Aaronson. Available online: http://www.scottaaronson.com/blog/?p=1823

Adriaans, P. (2019). “Information”, The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.)

Awodey, S. (2010). “Category Theory”. Oxford University Press.

Baez, J.C., Stay, M., (2009). “Physics, Topology Logic and Computation: A Rosetta Stone.”

Balduzzi, D. & Tononi, G. (2009). “Qualia: the geometry of integrated information.” PLoS computational biology, 5(8), e1000462.

Barrett, A. B. (2014). “An integration of integrated information theory with fundamental physics.” Frontiers in psychology, 5, 63.

Bateson, G. (1972). “Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.” University of Chicago Press.

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

Carnap, R. (1950). “Empiricism, semantics, and ontology”. Revue Internationale de Philosophie 4 (11):20–40.

Casali, A. G., O. Gosseries, M. Rosanova, M. Boly, S. Sarasso, K. R. Casali, S. Casarotto, M.-A. Bruno, S. Laureys, G. Tononi and M. Massimini (2013). “A theoretically based index

of consciousness independent of sensory processing and behavior." Sci Transl Med. 5: 198ra105–198ra105.

Chalmers, D. J. (1995). "Facing up to the problem of consciousness." Journal of Consciousness Studies 2 (3):200-19.

Chalmers, David J. (1996). "The Conscious Mind: In Search of a Fundamental Theory." Oxford University Press.

Chalmers, D. J. (2006). "Phenomenal Concepts and the Explanatory Gap". Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford: Oxford University Press.

Chalmers, D. J. (2009). "Ontological anti-realism". In David John Chalmers, David Manley & Ryan Wasserman (eds.), Metametaphysics: New Essays on the Foundations of Ontology. Oxford University Press.

Dehaene, S. & Changeux, J.-P. (2011). "Experimental and theoretical approaches to conscious processing." Neuron 70(2): 200–227.

Frankish, K. (2016). "Illusionism as a Theory of Consciousness." Journal of Consciousness Studies 23 (11-12):11-39.

Gazzaniga, M. S. (1967). "The split brain in man." Scientific American, 217(2), 24-29.

Grasso, M. (2019). "IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience." Journal of Consciousness Studies 26 (1-2):48-75.

Goff, Philip (2016). The Phenomenal Bonding Solution to the Combination Problem. In L. Jaskolla (ed.), Panpsychism: Contemporary Perspectives. Oxford University Press.

Goff, P. (2019) "Galileo's Error: Foundations for a New Science of Consciousness". Pantheon Books

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). "Quantifying causal emergence shows that macro can beat micro." Proceedings of the National Academy of Sciences, 110(49), 19790-19795.

Hume, David (1748/ 1999). "An Enquiry Concerning Human Understanding." Ed. T. L. Beauchamp. Oxford: Oxford University Press.

Jackson F. (1986), "What Mary Didn't Know", Journal of Philosophy, 83: 291–295

Kammerer, F.. (2016). "The hardest aspect of the illusion problem — And how to solve it." 23. 124-139.

Kripke, Saul (1980). "Naming and Necessity." Harvard University Press.

Lemon, R. N. & Edgley, S. A. (2010) "Life without a cerebellum." Brain 133, 652–654.

Levin, J. (2018) "Functionalism", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.)

Levine, J. (1983). "Materialism and qualia: the explanatory gap." Pacific Philosophical Quarterly, 64: 354-361.

Lombardi, O., & López, C. (2018). "What Does 'Information' Mean in Integrated Information Theory?." Entropy, 20(12), 894.

Marquis, Jean-Pierre (2019). "Category Theory", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.)

Masafumi O., Larissa A., & Tononi G. (2014) "From the Phenomenology to the Mechanisms of Consciousness: Integrated InformationTheory 3.0." PLoS: Computational Biology 8 (2014): e1003588.

Massimini, M., F. Ferrarelli, R. Huber, S. K. Esser, H. Singh and G. Tononi (2005). "Breakdown of cortical effective connectivity during sleep." Science 309: 2228–2232.

McQueen, K. J. (2019a). "Interpretation-Neutral Integrated Information Theory". Journal of Consciousness Studies 26 (1-2):76-106.

McQueen, K. J. (2019b). "Illusionist Integrated Information Theory." Journal of Consciousness Studies 26 (5-6):141-169.

Mindt, G. (2017). "The Problem with the 'Information' in Integrated Information Theory." Journal of Consciousness Studies 24 (7-8):130-154.

Moon, K., & Pae, H. (2019). "Making Sense of Consciousness as Integrated Information: Evolution and Issues of Integrated Information Theory." Journal of Cognitive Science, 20(1), 1-52.

Mørch, Hedda Hassel (2019a). "Is Consciousness Intrinsic?: A Problem for the Integrated Information Theory." Journal of Consciousness Studies 26 (1-2):133-162(30).

Mørch, Hedda Hassel (2019b). "The Argument for Panpsychism from Experience of Causation." In William Seager (ed.), The Routledge Handbook of Panpsychism. Routledge.

Nagel, T. (1974). "What is it like to be a bat?" Philosophical Review 83 (October):435-50. Oxford, UK.

Pautz, A. (2018). "What is Integrated Information Theory a Theory Of?"

Pearl, J. (2000). "Causality: Models, Reasoning and Inference"; Cambridge University Press: Cambridge, UK.

Popiel, N.J.; Khajehabdollahi, S.; Abeyasinghe, P.M.; Riganello, F.; Nichols, E.S.; Owen, A.M.; Soddu, A. (2020). "The Emergence of Integrated Information, Complexity, and 'Consciousness' at Criticality." Entropy 2020, 22, 339.

Russell, B. (1927). "The Analysis of Matter." London: Kegan Paul, Trench, Trubner & Co.

Schopenhauer, A. (1859). "The World as Will and Representation." Trans. E. F. J. Payne, vol. 1. New York: Dover.

Seager, William E. (1995). “Consciousness, information, and panpsychism.” Journal of Consciousness Studies 2 (3):272-88.

Seager, William E. (2006). “The 'intrinsic nature' argument for panpsychism.” Journal of Consciousness Studies 13 (10-11):129-145.

Searle, J. (1992), “The Rediscovery of the Mind” (Cambridge, MA: The MIT Press) pp. 97-98

Searle, J. (2013). “Can Information Theory Explain Consciousness? ” New York Review of Books 10 (2013): 53–58.

Shannon, C. E., (1948). “A Mathematical Theory of Communication”, Bell System Technical Journal, 27(3): 379–423 & 27(4): 623–656.

Shoemaker, S. (1982), “The Inverted Spectrum”, Journal of Philosophy, 79: 357-81.

Smart, J. J. C., (2019). “The Mind/Brain Identity Theory”, The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.)

Strawson, Galen (2006). “Realistic monism - why physicalism entails panpsychism.” Journal of Consciousness Studies 13 (10-11):3-31.

Tononi, G. (2004). “An Information Integration Theory of Consciousness”. BMC Neuroscience 5 (1): 42.

Tononi, G. (2012). “Integrated information theory of consciousness: An updated account.” Archives italiennes de biologie. 150. 56-90. 10.4449/aib.v149i5.1388.

Tononi, G. (2016). “Why Scott Should Stare at a Blank Wall and Reconsider (or, the Conscious Grid).”

Tononi, G. (2017). “Integrated information theory of consciousness: Some ontological considerations.” The Blackwell Companion to Consciousness, 621-633.

Tononi, G. & Koch C. (2015). "Consciousness: Here, There and Everywhere?" Philosophical Transactions of the Royal Society of London B: Biological Sciences 370 (1668), 20140167.

Tononi, G., Larissa A., & Masafumi O. (2014). "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." PLOS Computational Biology 10 (5), e1003588.

Tsuchiya N., Taguchi S., Saigo H. (2015). "Using category theory to assess the relationship between consciousness and integrated information theory." Neurosci Res. 2016;107:1-7. doi:10.1016/j.neures.2015.12.007

Tsuchiya, N. (2017). "What is it like to be a bat?"—a pathway to the answer from the integrated information theory. Philosophy Compass 12 (3):e12407.

Woodward, J. (2003). "Making Things Happen: A Theory of Causal Explanation"; Oxford University Press:

Van Fraassen, B. C. (1989). "Laws and Symmetry." Oxford University Press.

Van Fraassen, B. C. (1991). "Quantum Mechanics: An Empiricist View." Oxford University Press.