Facing Up to the Hard Problem of Consciousness as an Integrated Information Theorist

Final preprint of: https://link.springer.com/article/10.1007/s10699-020-09724-7

Please cite as:
Chis-Ciure, R., Ellia, F. Facing up to the Hard Problem of Consciousness as an Integrated Information Theorist. *Foundations of Science* (2021). https://doi.org/10.1007/s10699-020-09724-7

Abstract

In this paper we provide a philosophical analysis of the Hard Problem of consciousness and the implications of conceivability scenarios for current neuroscientific research. In particular, we focus on one of the most prominent neuroscientific theories of consciousness, Integrated Information Theory (IIT). After a brief introduction on IIT, we present Chalmers' original formulation and propose our own Layered View of the Hard Problem, showing how two separate issues can be distinguished. More specifically, we argue that it's possible to disentangle a Core Problem of Consciousness from a Layered Hard Problem, the latter being essentially connected to Chalmers' conceivability argument. We then assess the relation between the Hard Problem and IIT, showing how the theory resists conceivability scenarios, and how it is equipped to face up to the hard problem in its broadest acceptation.

Keywords: consciousness; Hard Problem; conceivability argument; Integrated Information Theory; physicalism; Russellian monism

1 Introduction

Consciousness[1] seems particularly hard to fit into our scientific worldview when we consider its subjective and qualitative aspect. Neurobiological theories that account for consciousness starting from neural mechanism seem unable to explain how physical matter gives rise to experience[2]: why certain neural processes are accompanied by certain experiential features, while others are not? Why seeing red gives *that* sensation, while pain a different one? This is the Hard Problem of consciousness (Chalmers 1995/2010), and it's generally assumed that any theory attempting to

[1] Both authors have contributed equally to the paper.
[2] In this article we use interchangeably the terms "consciousness", "experience", "phenomenal experience", "subjective experience", "phenomenality".

explain this feature of reality needs to address it in order to put the mind and the body "back together".

So far, no theory has been successful in solving this problem. Moreover, the Hard Problem appears more severe in the case of those theories attempting to squeeze consciousness out of some physical process that is proposed as explanatory for it. Such theories implement what we call a *mechanism-first* methodology: they start from some physical properties and try to infer and/or explain the phenomenology of experience. As one of the most prominent theories of consciousness, Integrated Information Theory (IIT) has gained attention, among others, for its mathematical model which in principle allows to characterize consciousness in both its quantitative and qualitative aspects. In contrast to other theories, IIT adopts a *phenomenology-first* approach, starting from the essential properties of experience, and trying to determine the physical properties that the physical substrate of consciousness should have (Tononi et al. 2016).

In this article we discuss how an integrated information theorist should address the Hard Problem of consciousness (HP). To do so, we introduce our own *Layered View of the Hard Problem*, which disentangles two central issues: a *Core Problem of Consciousness* (CPC), traditionally known as the mind-body problem, and the *Layered Hard Problem* (LHP), which arises when the core problem and Chalmers' conceivability argument are taken in conjunction. While the core problem reflects the fact that it's difficult to reconcile conscious experience and the physical world, the layered hard problem is the much stronger metaphysical thesis that doing so is possible only if they are essentially distinct entities. In the second section we introduce IIT, along with some ontological assumptions and implications of the theory. The third section provides an analysis of the problem as proposed by Chalmers, and presents our own Layered View of the Hard Problem. In section four we focus in more detail on Chalmers' conceivability argument and his two-dimensional semantics. In the fifth section we evaluate IIT under the light of the proposed analysis of the HP, and conclude that an integrated information theorist can give an account for the core problem, but should properly reject the two-dimensional conceivability argument and the layered problem that comes with it.

2 A Brief on Integrated Information Theory

Since the publication of the seminal paper *Towards a Neurobiological Theory of Consciousness* (Crick & Koch 1990), the neuroscientific research about consciousness has focused on the conceptualization and identification of the neural correlates of consciousness (NCC), namely the minimal set of joint neural mechanism and processes that are sufficient for consciousness. While progress has been made, no ultimate consensus on the NCC has been reached. Even more importantly, doubts have been raised whether or not this approach will eventually lead to a general theory of consciousness. In fact, NCC have been generally characterized as sufficient for consciousness (Crick & Koch 1990). However, sufficiency implies correlation, but not vice-versa (Ellia 2020).

In its various forms (Tononi 2004; Balduzzi & Tononi 2008; Oizumi et al. 2014; Tononi et al. 2016; Haun & Tononi 2019), Integrated Information Theory (IIT) presents itself as a theory of consciousness that can assess the quantity and quality of subjective experience in a physical system. "Understanding consciousness requires not only empirical studies of its neural correlates, but also a principled theoretical approach that can provide explanatory, inferential, and predictive power" (Oizumi et al. 2014). Such an approach is necessary since the neural and behavioral correlates of consciousness can be insufficient or misleading. Take, for example, cases where differences in behavior are insufficient to gauge the state of consciousness, such as a patient affected by unresponsive wakefulness syndrome and a patient in a locked-in state: despite very similar behavioral responses, we have good reasons to believe that they are different from their own intrinsic perspective. The unreliability of behavioral and neural correlates of consciousness is further remarked for systems progressively divergent from that of neurotypical adult human cerebral cortex. As the scale of alien-ness increases, the difficulty in determining the presence and character of consciousness in such systems increases exponentially. Consider the hypothetical scenario in which we establish a perfect correlation between consciousness and a certain pattern of activity in the neurotypical adult brain. How could this help us to determine whether or not an infant or a non-human mammal is conscious at a given time? And what to say about more controversial cases such as brain-injured patients, non-mammalian animals and perhaps even machines? In order to address these cases, pure correlations based on a sufficiency criterion are not enough and a principled approach which implies both sufficiency and necessity is needed instead (Tononi & Koch 2015).

In order to provide such a principled approach and to account for the level and quality of consciousness IIT employs a different strategy: rather than starting bottom-up from neural mechanism to consciousness, IIT adopts a *phenomenology-first approach*, namely from phenomenology to the mechanisms of consciousness. As such, the theory starts with five *axioms* derived from introspection which characterize the essential properties of every subjective experience. IIT defines axioms as "self-evident truths" (Oizumi et al. 2014); self-evidentiality here is not intended as immediately evident to anyone, but as necessary true upon reflection. These axioms are: Intrinsicality, Composition, Information, Integration and Exclusion. According to them, a conscious experience: exists intrinsically for the subject that it's experiencing it, i.e. it cannot be experienced by an external observer (Intrinsicality); it's structured in the sense that is composed by phenomenal distinctions bounded by relations (Composition); it's informative, in the sense that every experience, being the way it's, necessarily differs from the repertoire of other possible experiences (Information); it's integrated, in the sense that is not reducible to any proper subset of the phenomenal distinctions that compose it (Integration); and it's exclusive, in the sense that the content and the spatio-temporal grain of an experience are definite, it contains what it contains, nothing more or less, and there are no superpositions (Exclusion) (Oizumi et al. 2014; Tononi 2015; Tononi et al. 2016; Haun & Tononi 2019).

To each axiom corresponds a *postulate*. Postulates are inferred from axioms and they are defined as propositions that capture the ontological properties of the *physical substrate of*

consciousness (PSC). It's important to notice that IIT endorses the Eleatic principle as ontological criterion (Tononi 2015; Grasso 2019), meaning that according to the postulates, the PSC must have cause and effect power upon itself[3], namely must have the casual power to affect and be affected by itself (Intrinsicality). Furthermore, the PSC must have a causal structure composed by distinctions bounded by relations (Composition); the causal structure of the PSC must be specific, i.e. informative (Information); and irreducible to the causes and effects of its subcomponents (Integration). Finally, the PSC causal structure must be definite (i.e. its distinctions and relations are fixed and not in a superposition) and it must trump every other overlapping or partially overlapping causal structure at any spatial and temporal grain (Exclusion). The postulates can be translated into formal language, thus providing the means for a measure of consciousness in terms of integrated information (Φ), and a mathematical description of the Cause-Effect Structure (CES) in terms of information-geometry[4] (Albantakis 2017).

Alongside these five axioms and postulates, IIT posits an *identity* between the phenomenological properties of experience and the CES of a system:

> "The maximally irreducible conceptual structure[5] (MICS) generated by a complex of elements is identical to its experience. The constellation of concepts of the MICS completely specifies the quality of the experience (its quale '*sensu lato*' (in the broad sense of the term)). Its irreducibility Φ^{Max} specifies its quantity. The maximally irreducible cause-effect repertoire (MICE) of each concept within a MICS specifies what the concept is about (what it contributes to the quality of the experience, i.e. its quale '*sensu stricto*' (in the narrow sense of the term)), while its value of irreducibility φ^{Max} specifies how much the concept is present in the experience." (Oizumi et al. 2014).

Therefore, IIT accounts for consciousness by positing an identity between an experience and a *Cause-Effect Structure* that is *maximally irreducible* (MICS) (Tononi 2015). The nature of this identity is metaphysical, meaning that according to IIT the intrinsic experience of a system is ontologically identical with its Cause-Effect Structure. Moreover, we can take this identity in an explanatory sense as the blueprint to make inferences about the world, including other systems which have an intrinsic point of view. Therefore, the identity serves as an epistemic medium to investigate other consciousnesses from without, despite the fact that our own phenomenal access is the foundation of all our knowledge and limited to ourselves. After unfolding a system, if a maximally irreducible Cause-Effect Structure is found, an observer is able from the extrinsic perspective to entertain the claim that the system is conscious. Moreover, since cause-effect structures are quantified by Φ, its value will account for the quantity or level of consciousness

[3] According to the Eleatic Principle, to exist means to have causal power. IIT further refines this ontological criterion by requiring both cause *and* effect power. Moreover, according to IIT's ontology, in order to exist intrinsically, an entity must have maximally irreducible cause-effect power upon itself (Tononi 2015).
[4] For a complete description of IIT mathematical model, see Oizumi et al. (2014), Mayner et al. (2018).
[5] Earlier versions of IIT use the term "conceptual structure" instead of Cause-Effect Structure. This nominal change does not affect our argument in any way. The same applies with the replacement of "concepts" by "distinctions".

present in the candidate system, while the way in which the distinctions that compose the CES are related will describe its qualitative character of consciousness.

An important point is that this identity is not strictly between consciousness and its physical substrate: IIT identifies it with the Cause-Effect Structure that is *specified* by a system's complex of elements in a state. The physical configuration and dynamics of elements in a complex specify a CES that *is* an experience: the properties of the experience are identical to the properties of the relevant CES. In other words, consciousness *is* how an integrated system exerts cause-effect power upon itself (or intrinsically), independent from an extrinsic observer. Finally, while IIT's formalism allows for a mathematical description of the CES, the structure itself is a physical object rather than an abstract or mathematical one, in virtue of the fact that the physical substrate of consciousness does not exist generically, but exists *as* a Cause-Effect Structure. "[W]hatever the set of micro-elements that ultimately constitute the relevant macro-units and thereby the PSC, it's the [cause-effect] structure specified by the PSC at a particular moment that is identical to an experience, not the PSC taken simply as a set of connected units in a state" (Tononi 2017, 250).

IIT makes multiple predictions about consciousness and its underlying neural mechanisms. Among others, it predicts that the global maxima of Integrated Information in a neurotypical human brain are located in the posterior hot zone (Koch et al 2016; Tononi et al 2016) and that consciousness rises and falls accordingly to the presence or absence of connectivity of physical pathways (Tononi 2015). One of the most remarkable empirical successes of IIT was the development of the Perturbational Complexity Index (PCI), a proxy measurement for Φ (Casali et al. 2013), sensitive enough to distinguish between different patient states[6]. To summarize, IIT is a neuroscientific theory of consciousness which takes the connection between a physical system in a state and the character of its phenomenal experience as *explanandum* and the phenomenology-first approach as *explanans*; moreover, by making predictions which can be empirically tested (at least in principle), IIT is falsifiable (Tononi et al 2016; Tsuchiya et al. 2020).

3 The Layered View of the Hard Problem

The joint endeavor of philosophy and cognitive sciences to explain this most intimate and yet elusive phenomenon of consciousness has been permeated by a methodological distinction between *easy problems* and the *Hard Problem* of consciousness (Chalmers 1995/2010). This distinction can be *prima facie* understood as a difference in the explanations needed to account for their respective *explananda*. On one hand, the easy problems are vulnerable to explanations in terms of structural configuration and functioning in physical systems, the kind of explanations

[6] PCI is an algorithm based on Transcranial Magnetic Stimulation and High Density-EEG data which allows for reliable predictions with respect to the brain's capacity for sustaining phenomenal experience, in absence of behavioral responses, including reports. Notably, PCI has proved effective in discriminating between different states: wakefulness, dreamless sleep, dreaming, and anesthesia under different agents, such as ketamine and propofol. This is important because these states are characterized by different levels of consciousness and substantial dissimilarities in the richness and vividness of conscious contents (Massimini et al. 2010; Casali et al. 2013; Sarasso et al. 2015).

obtained via the methods of natural sciences, including cognitive neuroscience. There is nothing more required to explain such problems than to specify a computational and/or neural mechanism that performs the relevant function. That is why they are called "easy problems". This type of explanation is called physical, since it implies an account of the *explanandum* in terms about physical processes. On the other hand, Chalmers argues, the HP of consciousness is resistant to such methods, requiring instead a *non-reductive* explanation, where consciousness itself is taken as fundamental, i.e. not explainable in simpler terms.

But what are the easy problems and the HP? According to Chalmers (1995/2010; 1996; 2009/2010; 2018), the easy problems are those of explaining various mental functions, like attention, perceptual integration, conscious access, reportability, memory, and others[7]. In contrast, the HP is that of explaining *what it is like to be us* (Nagel 1974), i.e. phenomenal consciousness (or subjective experience). For any system endowed with consciousness, *there is something it is like to be* that system ("creature consciousness" or "intransitive consciousness"); accordingly, for any conscious mental state, *there is something it is like to be in* that state ("state consciousness" or "transitive consciousness") (Gennaro 2019). This phenomenal or subjective character constitutes the target of the HP. Chalmers holds that specifying mechanisms that play some functional or causal role within a given conscious system is not sufficient to explain subjectivity or experience, but it's sufficient for the phenomena of the easy problems.

3.1 The Core Problem of Consciousness (CPC)

What justifies the distinction between the explanations required? Is there a reason why phenomenal consciousness cannot be explained in terms of physical processes? This cannot just be accepted dogmatically without justification. As Chalmers (1995/2010) puts it, it's a *conceptual fact* that mechanistic explanations of physical sciences are insufficient to explain experience, but are adequate to account for the easy problems. For the latter all it *could possibly* be asked for is a specification of the physical mechanisms responsible for the relevant functions or causal roles. To avoid pushing the dogmatic stance a level up or down, a further justification is needed. The justification for this conceptual fact is given by another conceptual fact: the *conceptual coherency* of a scenario where, given any physical process, it *could* be instantiated in the absence of experience. In principle, we could *conceive* of any physical process that is put forward as the basis of consciousness as being instantiated without any phenomenal aspect at all[8]. This conceptual coherency fact justifies the claim that the HP is impregnable to methods employed to solve the easy problems.

If the preceding analysis is correct, then the HP has a layered structure. We propose the *Layered View of the Hard Problem*, according to which there is a *core problem* and a further *conceptual*

[7] For more details on the easy problems, see Chalmers (1995/2010, 3-6).

[8] We suspect that the conceivability argument evolved from Chalmers' dissatisfaction with early models of consciousness that seemed *ad hoc* or gave an impression of *fiat* argumentation.

layer that together constitute the HP. The first layer is captured by the fact that there is something it is like to be us (Nagel 1974). Call this the *Core Problem of Consciousness* (CPC).

> **(CPC)** We need to explain how the fact that there is something it is like to be us relates to physical matter.

CPC makes experience an *explanandum* in its own right. This fact requires no extra justification, since it's something we are directly acquainted to, and almost everybody takes experience as real and in need of scientific explanation. Furthermore, since at least the dawn of Modernity it has been widely recognized that, although there is a consistent connection between consciousness and physical matter (at least in the form of biological bodies and brains), the two seem impermeable to reconciliation and unification in a single theoretical framework. In fact, one can read Descartes' *Second Meditation* (1641/1948) as offering a conceivability argument for the real distinction between mind and body, based on the fact that I can conceive of my mind existing without my body, but I cannot coherently conceive of my body existing without itself. Thus, via the principle of the distinctness of discernibles, my mind is not identical to my body[9]. This is just a quick illustration of the fact that the Core Problem of Consciousness (CPC) is actually the mind-body problem that contemporary philosophy and science inherited from our intellectual tradition, expressed in the current parlance of what-is-it-likeness. CPC arises due to the existence of phenomenality coupled with its stark incompatibility or incommensurability with the physical world of atoms, neurons, and bodies. Solving CPC is one of the greatest challenges that mankind has to face in its pursuit of knowledge.

3.2 Layered Hard Problem (LHP)

We don't equate what we call the "Core Problem of Consciousness" with what Chalmers' dubbed the "Hard Problem". Therefore, we disagree that "to generate the hard problem of consciousness, all we need is the basic fact that there is something it is like to be us" (Chalmers 2018, 49-50). The fact of experience or subjectivity is sufficient *only* for CPC, provided that its connection with physicality is taken into account. Chalmers' HP requires one further step. What we claim is that even though the relation between phenomenality and physicality has been considered problematic for a long time, it does *not* by this fact create the HP as Chalmers proposes it. HP is a specific instance of problematization which relies on extensive use of conceivability and modal concepts. The conceptual layer is given by the fact that conceivability scenarios are coherent, which provides

[9] Not to mention the famous *Sixth Meditation*, where Descartes argues for a *distinctio realis* between minds and bodies by an appeal to a difference in essential properties or nature. Thus, minds and bodies fail to share all properties required by an alleged identity between them, and the well-known *locus* of difference is the property of extendedness and thus of divisibility. Since only bodies are extended, then, by distinctness of discernibles, the mind is distinct from the body.

justification for the fact that experience cannot be explained in terms of structure and function. We call this the *Layered Hard Problem* (LHP)[10].

> **(LHP**) There is something it is like to be us and we need to explain this fact. Since conceivability scenarios are coherent, a mechanistic explanation in terms of physical processes is insufficient for this; therefore, we need an alternative explanation.

Obviously, LHP contains CPC, but adds further *epistemic claims* about how an explanation should look like. The HP of "why does physical processing give rise to experience at all?" requires this layered view.

> "For any physical process we specify there will be an unanswered question: why should this process give rise to experience? Given any such process, it's conceptually coherent that it could be instantiated in the absence of experience. It follows that no mere account of the physical process will tell us why experience arises. The emergence of experience goes beyond what can be derived from physical theory." (Chalmers 1995/2010, 14)

To repeat, we are far from claiming that, given the present the analysis, the relation between phenomenality and physicality is not highly problematic – or that it's not a "hard" task for any theoretical account. What we claim is that Chalmers' HP is a specific way of cashing out this inherent difficulty already expressed in the traditional mind-body problem and our CPC, by superposing a conceptual layer given by conceivability considerations and the modal apparatus of possible worlds and the intensions of terms like "physical" and "phenomenal" at those worlds, which we briefly present in the next section. Construed like this, HP is part of a more comprehensive *corpus* of metaphysical (e.g. zombies, inverts) and epistemological (knowledge, epistemic asymmetry) arguments[11] against a materialist or physicalist theory of consciousness. However, as our analysis shows, the problem itself is generated via argumentative mechanisms like the conceivability of zombie worlds. The latter correspond to those conceivable scenarios where you have a perfect duplicate or "replica" of a physical process putatively explanatorily relevant for experience, yet without any phenomenality or consciousness.

> "[*Z*]*ombie*: a system that is physically identical to a conscious being but that lacks consciousness entirely. […] [T]heir brain processes will be molecule-for-molecule identical with the original, and their behavior will be indistinguishable. But things will be different from the first-person point of view." (Chalmers 2003/2010, 106-7)

[10] The standard Hard Problem is actually the Layered Hard Problem (HP=LHP). From now on, when we mention any term of the two, you can replace it by the other, so they are considered synonyms.
[11] See Chalmers (1996; 2003/2010). Here we focus exclusively on the conceivability argument.

Put simply, if *p* is a physical (brain) process which a given theory takes to *be* or *be essential for* a given experience *q*, then we could coherently conceive of a scenario in which *p* obtains, yet *q* does not. This is meant to imply that for an arbitrary physical process *p*, and an arbitrary phenomenal experience *q*, *p* is (metaphysically) distinct from *q*, i.e. $p \neq q$. Let's have a closer look at this argument.

4 The Two-Dimensional Conceivability Argument

Materialism (or physicalism[12]) states that everything is material (or physical). In its crudest form, materialism about consciousness entails that consciousness is a material entity. By the necessity of identity, materialism is taken to be a modal thesis: in every possible world in which there is phenomenal experience, it *could not have been the case* that it's non-material. If consciousness is material, it's *necessarily* so. The modality at stake here is *metaphysical modality*. The conceptual layer, i.e. the coherency of conceivability scenarios, renders any materialistic account of experience impotent in the face of HP. Basically, almost[13] all physicalist theories of consciousness are threatened by the Conceivability or Zombie Argument[14] (Chalmers 1996; 2010, ch. 6). In the two-dimensional framework used to articulate the refined version of the argument, three related distinctions are important: (i) between primary and secondary conceivability; (ii) between primary and secondary possibility; and (iii) between primary and secondary intensions of an expression. We introduce them briefly.

4.1 Primary and secondary: conceivability, possibility, and intensions

Primary conceivability. In general, a possible world is an alternative history or state of the world, a way in which our world could have turned out (Kripke 1980). In particular, a possible world *w* is primary (or 1-) conceivable when one conceives *w* as being qualitatively identical to the actual world. 1-conceivability is tied to the rational or *a priori* domain: a 1-conceivable scenario cannot be ruled out a priori, so it's (at least) logically coherent. For instance, I cannot coherently conceive of a world where circles have four edges, or of one that is not identical with itself. For a world or scenario to be primary (or 1-) conceivable, it must respect the cannons of logical intelligibility. In this sense, bracketing any empirical considerations, we can say that "water is not H_2O" is 1-conceivable, so the identity between water and H_2O is not established in a relevant sense by logic alone. Chalmers' talk of scenarios is in terms of centered possible worlds, constituted by a rational agent, the world, and a specific time. The main idea is that there is an individual at a time at the centre of the scenario engaged in rational reflection, and we consider this hypothetical scenario *as*

[12] We use "materialism" and "physicalism" as synonyms, albeit they can be distinguished (Bunge 2010).
[13] With the exception of Russellian monism (RM), if it's truly a version of physicalism about experience.
[14] Similar considerations apply for "partial zombies" and/or "(partial) inverts".

actual, which means that we treat it like it would be our own world, and we evaluate what is rational to endorse.

Secondary conceivability. When we consider a possible world *w* as secondary (or 2-) conceivable, we evaluate it relative to how the actual world is. If in 1-conceivability we take the scenario *as if it would be the actual world* and ask how a rational agent would judge a statement based on the evidence available, in 2-conceivability we take the possible world *as counterfactual*, namely as a way in which our own world could have turned out. *A posteriori* considerations make it the case that water=H_2O. Accepting that natural kinds are rigid designators (Kripke 1980), so the identity holds across all possible worlds, it's not the case that "water is not H_2O" is 2-conceivable. Because said identity holds by metaphysical necessity, "water is not H_2O" is only 1-conceivable, and not 2-conceivable. For a given statement *S*, its credentials dictate its conceivability status: if it's a priori coherent, then it's 1-conceivable; if it's an *a posteriori* true identity, then it's also 2-conceivable. To exemplify, say I conceive of a world where water is not H_2O, but XYZ. Yet in the actual world it's true that water=H_2O. Therefore, water is necessarily the same thing as H_2O (by necessity of identity). This means that the world I conceived, where water was XYZ, is not 2-conceivable, because it violates a truth of our world, namely that water=H_2O. Once again, in 2-conceivability we evaluate possible worlds relative to the actual world. Roughly, primary conceivability is regulated by logic and armchair reasoning, whereas secondary conceivability depends on the nomological profile of our world[15].

Primary and secondary possibility. Simply enough, if a scenario is 1-conceivable, then it's 1-possible. If it's also 2-conceivable, then it's also 2-possible. Primary conceivability is a guide to primary or *epistemic* possibility, while secondary conceivability is a guide to secondary or *metaphysical* possibility. Quite obviously, 1-possibility is established *a priori*, whereas 2-possibility depends on *a posteriori* matters. A scenario *s* can be 1-conceivable, hence 1-possible, but 2-inconceivable, hence 2-impossible. "Water is not H_2O" is such a case. The crucial problem with conceivability and possibility is easy to comprehend: 1-conceivability entails 1-possibility, but *not* 2-possibility. So even if the proposition "consciousness is distinct from matter" is logically coherent, i.e. it's 1-possible or epistemically possible, that doesn't imply immediately that it's also a metaphysical truth, i.e. that it's 2-possible or metaphysically possible. Chalmers needs 2-conceivability and 2-possibility for his argument to go through, which he cannot establish by mere logical considerations and *a priori* reflection. This brings us to the problem of primary and secondary intensions.

Intensions. The conceivability argument gets its philosophical nuance from Chalmers' treatment of intensions in his two-dimensional semantics, which is a version of possible world semantics. In possible worlds semantics, an expression has both an extension and an intension. The extension of a sentence *S* is its truth-value, while its intension is a function from possible worlds to extensions, namely the intension of *S* at a possible world is true only if it's the case that *S* in that world (e.g. the proposition "Socrates was sentenced to death" is true at a possible world only if Socrates suffered this fate there). The extension of a general term like "brains" is the class

[15] We focus solely on ideal conceivability, leaving aside prima facie conceivability.

of objects denoted by the term (i.e. brains), while the intension of a term is again a function that maps a possible world to its extension: the set of all brains in a possible world is the intension of "brain" at that world. So, if I conceive of a possible world with brains floating in vats and connected to a Matrix, the intension of the term "brain" is simply the whole set of brains at that world, i.e. its extension. In short, the intension of an expression (e.g. a singular or general term, or a sentence) is a function from possible worlds to extensions.

Primary intension. Chalmers introduces a distinction between primary and secondary intensions, bearing on his distinction between (centered) possible worlds considered as actual and possible worlds considered as counterfactual. 1-intensions have to do with the former, while 2-intensions are connected to the latter. The primary (or 1-) intension of a term *t* is given by whatever determines the extension of that term in a scenario (centered possible world) considered as epistemic possibility (or 1-possible); the 1-intension of a proposition at a centered world is its truth-value at that world considered as actual. According to Chalmers (2010, 150), the 1-intension of a term is tied to a certain theoretical (causal or functional) role it plays in our cognitive economy. The 1-intension of "proton" picks out whatever property or bundle thereof plays proton's theoretical role (e.g. being constitutive of atomic nuclei). A world *w* in which there is something that plays the causal or functional role of a term *t* is said to *verify t*. At that world, *t* has a 1-intension, so it's 1-conceivable, hence 1-possible.

Secondary intension. The secondary (or 2-) intension of a term *t* is tied to the role it actually plays in our world with its actual laws. The 2-intension is a function from the actual world to possible worlds considered as counterfactual: if protons, given their nomological profile, play the role of what is picked out by the term "proton" in our world, then the 2-intension of "proton" in every other world will pick out either protons, or nothing. A world *w* in which there is something that actually plays the causal or functional role of a term *t*, is said to *satisfy t*. A possible world *w* might verify a term *t*, yet fail to satisfy it. A possible world in which there isn't anything that plays the role of protons and in which there are no protons does not verify or satisfy "proton". To contrast, for 1-intensions it's enough that there is something that *plays the role* the referent of the term does, so it's not necessary that they are the same (class of) objects. If protons and schprotons (fictitious objects) play the same role, all worlds containing schprotons will verify "proton". However, the 2-secondary intension of a term is tied to the role it *actually* plays in our world. So, a possible world satisfies "proton" iff there are protons there. If there are only things that do the same job (e.g. schprotons), that world does not satisfy "proton", it only verifies it. This applies *mutatis mutandis* to statements. For a statement *S*, there will be some possible worlds that only verify but not satisfy it. Arguably, for all *a posteriori* Kripkean necessities, there will be some possible worlds verifying, but not satisfying them. "Water is H_2O" is verified *and* satisfied by those possible worlds where water is H_2O, but it's *only* verified and *not* satisfied by those worlds where there is a colorless drinkable liquid which it's not H_2O, but XYZ.

4.2 The 2D Conceivability Argument

Thus, to finally get to Chalmers' view, consider *P* a conjunction of all actual microphysical truths (e.g. the set of all truths of fundamental physics), and *Q* an arbitrary phenomenal truth (e.g. system *x* has phenomenal consciousness). As Chalmers (2010, 152) puts it, the refined two-dimensional conceivability argument goes like this:

(1) *P*&~*Q* is conceivable.

(2) If *P*&~*Q* is conceivable, then *P*&~*Q* is 1-possible.

(3) If *P*&~*Q* is 1-possible, then *P*&~*Q* is 2-possible or Russellian monism is true.

(4) If *P*&~*Q* is 2-possible, then materialism is false.

(C) Materialism is false or Russellian monism is true.

For the argument to go through, i.e. for the step from 1-possibility to 2-possibility to be feasible, both the 1- and 2-intensions of *P* and *Q* must coincide. Following Kripke (1980), Chalmers seems to accept as uncontroversial the epistemic transparency of phenomenal terms. In their case, appearance is being, in the sense that a phenomenal term referring to a qualitative content also reveals its essence, so *Q*'s 1- and 2- intensions coincide. If something appears to be pain, is sufficient to claim it is pain.

Russellian Monism (RM) escapes the thrust of the argument because the 1-intension and 2-intension of physical terms in *P* can fail to coincide at some worlds. If one accepts as real the distinction between *intrinsic* or *categorical* properties on one hand, and *extrinsic* or *dispositional* properties on the other, then one can have a possible world *w* that is structurally identical to the actual world (with physics describing structural relations between entities), which however differs in the intrinsic profile (with intrinsic properties seen as the "categorical base" of structural properties). Such a world *w* with same structural (or extrinsic) and different intrinsic profile would verify a given physical term *t* in *P*, but not satisfy it. If identity in structural profile is insufficient to metaphysically necessitate any phenomenal truth in *Q*, premise 3 leaves untouched a view like RM, which requires that two worlds share both extrinsic and intrinsic profile for the same phenomenal truths to obtain in them.

In other words, zombies bite any physicalist theory that does not accept the distinction between intrinsic and dispositional properties, with the former "giving the categorical base" (Chalmers 2010, 151) for the latter, and phenomenal properties being identical with or being essentially related to intrinsic properties. To summarize, materialism about consciousness is vulnerable to the conceivability argument because the 1- and 2-intensions of physical terms can fail to coincide at some (centered) possible worlds, while the 1- and 2-intension of phenomenal terms must coincide[16], because for something to be a phenomenal content (e.g. pain) it's sufficient to appear as such (e.g. painful).

5 Integrated Information Theory and the Hard Problem

[16] The case where the 1- and 2-intension of phenomenal terms can differ can be dealt from IIT's perspective in the same way as the standard case presented above, so we don't discuss it further.

After introducing both IIT and HP, we can now address their relation and see what IIT can say about both CPC and LHP. We believe that any complete neuroscientific theory of consciousness should address the problem of experience, and the underlying metaphysical relation between it and its physical substrate, or at least present an explanatory apparatus which is compatible with such a metaphysical description. Considered on its own, IIT does so in an explicit way. In fact, one of the advantages IIT has over rival positions is the clarification of its assumptions, by means of the axioms, postulates and the identity. The identity is indeed the key to address both CPC and LHP. Elsewhere we argue how, given the peculiar phenomenology-first approach of the theory, the identity posited by IIT is by all means an a priori identity (Ellia, *in preparation*), and more specifically, it can be understood as a constitutive a priori identity (Chis-Ciure, *in preparation*).

First, we can say that the HP does not arise for a theory like IIT. Tononi (2015) seems to hold that this is an obstacle only for 'bottom-up' theories, i.e. those attempting to infer the existence of consciousness from physical processes. In contrast, IIT reverses the order: by capturing the essential properties of experience in its axioms, the theory infers the postulates describing the physical properties that a system must exhibit in order to be conscious. In this sense, IIT does not directly address LHP because there is no such problem for a theory of its type. However, one could also argue that to account indirectly for experience, which is taken as explanatorily primitive or fundamental, is a viable strategy for a theory, even though it's not further reducible to something else. Thus, a first conclusion is that IIT's own epistemology denies that there is any LHP to begin with; however, by reconciling experience with its physical substrate (i.e. through the identity between CES and experience), the theory addresses CPC. No matter whether IIT does so properly or not, it shows that our layered view on the hard problem is correct: it's possible to disentangle the core of the problem (CPC) from its justification provided by the conceivability argument (LHP).

More can be said about the relation between the integrated information view of consciousness and LHP when considering conceivability scenarios. Recall that LHP's conceptual layer presupposes the conceivability of "philosophical" or "molecule-for-molecule" zombies. There is an important difference between "philosophical" and "functional zombies"[17] (Oizumi et al. 2014, Tononi 2015). One of the corollaries of IIT is that sheer functional complexity does not entail consciousness. There are sophisticated, yet unconscious systems (e.g. those displaying a feed-forward architecture) that can perform functions identical with those performed by a system having high integrated information, provided enough number of units and time. This marks the difference in zombies: functional zombies are not physically identical to their counterparts, only functionally identical. However, "philosophical" zombies need *both* structural and functional identity – they are perfect physical replicas. Tononi (2015) is explicit: "if the postulated identity […] is true, a system of elements in a state that specifies [a] conceptual[18] structure has the corresponding experience *necessarily* and cannot be a zombie". If the argument is sound and the identity true,

[17] In the literature, these are also known as "perfect" or "true" zombies. We believe that "functional zombies" help the reader by clarifying what we mean by that.
[18] Like noted above, the term "conceptual structure" is synonymous to "cause-effect structure".

then a physical system in a state specifying a maximally irreducible cause-effect structure has the corresponding conscious experiences *necessarily*. Then, if further pressed, the integrated information theorist would reject the conceivability step: according to IIT, "philosophical" zombies are not conceivable, neither in the primary nor in the secondary sense, only "functional" ones are. Notice how this differs from the Kripkean cases of *a posteriori* necessary propositions like 'water is H_2O'. For Kripke, it is 1-conceivable that water is not H_2O, but it's not 2-conceivable for the reasons presented in section 4. On the contrary, the identity in IIT should be understood as a priori: by definition, two identical physical system will be phenomenally identical, as according to the theory there is no ontological distinction between the two[19]. This is of the utmost importance, because otherwise philosophical zombies could be at least 1-conceivable. Instead, given Chalmers' definition of "ideal conceivability", if one endorses IIT, then one cannot retain zombies as conceivable by definition. Thus, a second conclusion is that, by understanding IIT's identity properly, the theory directly denies the conceivability of zombie scenarios, thereby denying the justification for the second layer, thus LHP itself.

A more general refutation of LHP would take nomic or natural possibility as the proper modality of scientific inquiry rather than the metaphysical or logical one, and therefore reject the conceivability argument altogether, a move we suggest in the next subsection.

Finally, it is worth mentioning that IIT does not simply address CPC, but it does so in an exhaustive way. By explicating the axioms, the postulates, and the identity, the theory defines the essential properties of experience and how they relate to physical mechanisms. This ultimately provides specific answers to everlasting questions such as: (i) how to find the presence of consciousness within a system – by displaying a value of Φ greater than 0; (ii) how a manifold of input is experienced as a whole – by being a maximally integrated cause-effect structure; and (iii) why certain experiential features are bound together (i.e. the binding problem) – by the relations within a cause-effect structure.

5.1 Scientific Inquiry and the Relevance of Nomological Modality

We discussed so far how IIT rejects the conceivability argument from *within*, appealing to the conceptual apparatus proposed by the theory. In this subsection we provide a general rationale for rejecting Chalmers' conceivability argument within the context of scientific inquiry.

Chalmers relies on a monistic view of modality, modal rationalism, according to which modality is conceptually tied to the rational domain (rational consistency, apriority, conceivability). This modality is called logical modality: a statement *S* is logically necessary when *S* is a priori true. We raise two questions here. What is the relevance of modality for the scientific inquiry into the nature of consciousness? And how come a modality derived from the rational

[19] Notice how, according to our reconstruction, in IIT consciousness is identical to a physical object, the cause-effect structure. However, this is radically different from standard physicalism or other kinds of reductionism. Consciousness is ontologically identical but epistemically irreducible to physical stuff, as consciousness is our starting point. Moreover, 'physical' in IIT is defined within the theory as 'cause-effect power'.

domain can have substantial ontological significance? With regard to the first one, Chalmers claims that the space of logically possible worlds has a *heuristic value* for the scientific enterprise, in the sense that counterfactual thinking, relying on a logical modality concept, is a helpful tool for new discoveries. Even if this is so, these considerations are powerless against materialism about experience. When considering the answer to the second question, Chalmers (2010, 191) claims that it's "obvious" that such modal notions have a "bearing" on the ontological domain. The first example he gives is that of the a priori entailment from unmarried men to bachelors, which can thus be discarded from our ontology. However, this is a very weak reply, insofar 'Not all unmarried men are bachelors' is at best *prima facie* conceivable and only if one is not certain what the terms mean, or thinking about logical relations is not her strongest point. The second example is the modal status of materialism itself, "so the analysis of modality quite reasonably drives conclusions about materialism".

However, most people would agree that with regard to scientific discourse and practice, the only modality that has ontological significance is nomological or natural modality[20]. Admittedly, the logical modality tied to the rational domain can have heuristic value, but this is not substantive insofar ontological revisions derived from actual science are concerned. For this purpose, it's relevant only how our *actual world*, with its *properties* and *laws*, is. Logical possibility is ubiquitous in mathematics and logics, but only indirectly relevant in science (Bunge 1977; 2010; Mahner 2017). Scientists consider alternatives to the actual state of things that are bounded by the nomological profile of our universe, not outlandish scenarios regulated by logical consistency alone. It's obvious that everything that is naturally possible is logically possible, and everything that is logically impossible is naturally impossible. Nevertheless, not everything that is logically possible is naturally possible. Chalmers certainly does not entertain this, yet he uses the conclusion about materialism drawn by assuming this bearing of the logical on the ontological to argue for the relevance of the rational domain for ontology[21]. We believe that IIT authors implicitly endorse the same view about nomological modality and scientific inquiry:

> "As often happens in philosophical debate, every argument or thought experiment has a counter-argument or -experiment. Of course, zombies are no exception to this. This is an example of the circular traps that we would like to avoid, so we will take this argument no further. […] [W]e prefer to keep our feet on the ground, rather than experience the frisson of logical disorientation." (Massimini & Tononi 2018, 10)

[20] This has been developed with great rigor more than 40 years ago by Bunge (1977, ch. 4; 2010; also, Mahner 2017).
[21] Chalmers' considerations seem to be driven by his commitment to modal rationalism, which is a version of modal monism, namely the metaphysical thesis that there is only one type of necessity/possibility that governs all of our use of modal expressions. Without discussing any of these issue in detail, we want to point out to the reader that there are working alternatives to this view. For instance, Fine (2004) argues for a modal pluralistic view, where the metaphysical necessity (which roughly corresponds to Chalmers' broad notion of logical necessity) is irreducible to both natural or nomological necessity and normative necessity. The three types are to be considered as independent sources to which our modal expressions could be traced. Neither is analyzable in terms of the other.

For example, counterfactual reasoning is employed in science to evaluate causal relations beyond simple statistical correlations. In fact, perturbational data can be combined to obtain causal models of the target system. More specifically, IIT's own causal model relies on counterfactuals in assessing the joint constraints that the mechanisms of a system impose upon each other. Yet the modality at stake here is nomological: only those states of the system that are possible – in the sense of being alternative configurations of system's elements – are taken into consideration[22] (Albantakis et al. 2019; Albantakis et al. in preparation). However, accurate models require extensive perturbational data which are often difficult to obtain (Albantakis 2017).

6 Conclusion

In this article we have examined in detail the relation between the Hard Problem of Consciousness and Integrated Information Theory. We have introduced our Layered View on the Hard Problem in order to disentangle the Core Problem of Consciousness (CPC) from the conceivability argument, which constitutes a separate conceptual layer (LHP). We have shown IIT's stance on both the core and the layered problem. On one hand, IIT acknowledges CPC, and addresses it through its unique epistemology: the phenomenology-first approach. On the other hand, granted that one endorses IIT (i.e. accepts its axioms and postulates), the fundamental identity between a cause-effect structure and an experience entails that conceivability arguments are ungrounded, hence there is no LHP to solve to begin with.

We take LHP to be the version of the problem commonly accepted in the literature, i.e. the much stronger thesis that no theory can account for experience solely in mechanistic terms. We provide a way to reject this thesis, while preserving the validity of the "hard problem", i.e. CPC, as a genuine challenge for any proposed theory. Moreover, we believe we have done that by showing the epistemic advantage of neuroscientific theories that employ a phenomenology-first approach (such as IIT), over theories that emphasize neural mechanisms exclusively. Finally, we argued that scientific practice should focus its scope on nomological possibility instead of the much broader domain of logical possibility.

Acknowledgments We are thankful to Giulio Tononi, Matteo Grasso and Garrett Mindt for meaningful conversation about IIT and its philosophical foundations. We also thank Jussi Jylkkä, Henry Railo, Jarno Tuominen, and other members of Antti Revonsuo's *Consciousness Research Group* for helpful comments on a presentation we had at University of Turku, Finland. We are especially indebted to the audience of the *4th SILFS Postgraduate Conference*, University of Urbino, Italy, and to the kind organizers who invited us to submit the paper to this journal issue. We are also thankful to two anonymous reviewers for important feedbacks on an earlier version of the paper.

[22] Including those states that would not be observed without an intervention on the system.

Conflict of Interest Statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Albantakis, L. (2017). Integrated Information Theory. In *Beyond Neural Correlates of Consciousness*, Overgaard, M. et al. (eds.), Taylor & Francis Group, ProQuest Ebook.

Albantakis, L., Marshall, W., Hoel, E., Tononi, G. (2019). What Caused What? A quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, 21 (5), pp. 459.

Albantakis, L., Ellia, F., Tononi, G. (in preparation). A Causal Information Account of Actual Causation.

Balduzzi, D., & Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, 4(6), e1000091.

Bunge, M. (1977). *Treatise on Basic Philosophy. Volume 3: Ontology*. Boston: Reidel Publishing.

Bunge, M. (2010). *Mind and Matter a Philosophical Inquiry.* New York: Springer Science.

Casali, A., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med*. 5.

Chalmers, D. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D. (1995/2010). Facing Up to the Problem of Consciousness. In Chalmers, D., *The Character of Consciousness* (pp. 3-34). New York: Oxford University Press.

Chalmers, D. (2003/2010). Consciousness and Its Place in Nature. In Chalmers, D., *The Character of Consciousness* (pp. 103-139). New York: Oxford University Press.

Chalmers, D. (2009/2010). The Two-Dimensional Argument Against Materialism. In Chalmers, D., *The Character of Consciousness* (pp. 141-205). New York: Oxford University Press.

Chalmers, D. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies* 25(9-10), 6–61.

Chis-Ciure, R. (in preparation). The Central Identity of Integrated Information Theory as Constitutive A Priori.

Crick, F. and Koch, C. (1990). Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences*, 2.

Descartes, R. (1641/1984). Meditations on First Philosophy. In *The Philosophical Writings of Descartes*, vol. II, transl. by Cottingham, J., Stoothoff, R., Murdoch, D., Cambridge: Cambridge University Press.

Ellia, F. (2020). Francis Crick e il Problema Difficile della Coscienza. *Sistemi Intelligenti* 1/2020.

Ellia, F. (in preparation). Integrated Information Theory and the Axiomatic Approach.

Fine, K. (2004). The Varieties of Necessity. In *Conceivability and Possibility*, Gendler, T., Hawthorne, J. (eds.), Oxford: Oxford University Press.

Gennaro, R. Consciousness. The Internet Encyclopedia of *Philosophy*, ISSN 2161-0002, https://www.iep.utm.edu/consciou/#H7, Accessed 24 April 2019.

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

Koch, C. (2019). *The Feeling of Life Itself*. Cambridge: MIT Press.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17(5), 307-321.

Kripke, S., 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Levine, J. (1983). Materialism and phenomenal properties: the explanatory gap. *Pacific Philosophical Quarterly* 64 (4), 354–361.

Mahner M. (2017) The Philosophy of Mind needs a better Metaphysics. In Bunge, M. (2017) *Doing Science in the Light of Philosophy*. Singapore: World Scientific Publishing.

Massimini, M., F. Ferrarelli, M. J. Murphy, R. Huber, B. A. Riedner, S. Casarotto and Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive Neuroscience*.

Massimini, M., Tononi, G., (2018). *Sizing Up Consciousness: Towards an Objective Measure of the Capacity for Experience*, transl. by Anderson, F., New York: Oxford University Press.

Mayner, W., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., Tononi, G. (2018) PyPhi: A toolbox for integrated information theory. *PLoS Computational Biology*, 14(7), pp. e10061144.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435–450.

Northoff, G. Tsuchiya, N. Saigo, H. (2019) Mathematics and the Brain: A Category Theoretical Approach to Go Beyond the Neural Correlates of Consciousness, *Entropy*, 21(12), 1234.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5), e1003588.

Sarasso, S. Boly, M. Napolitani, M. Gosseries, O. Charland-Verville, V. Casarotto, S. Rosanova, M. Casali, A. Brichant, JF. Boveroux, P. Rex, S. Tononi, G. Laureys, S. Massimini, M. (2015) Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine, *Current Biology* 25.

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

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, 10(1), 4164.

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450-461.

Tononi, G. (2017). Integrated Information Theory of Consciousness: Some Ontological Considerations. In Schneider, S., Velmans, M (Eds.), *The Blackwell Companion to Consciousness* (pp. 621-633). Chichester: Wiley Blackwell.

Tsuchiya, N. Andrillon, T., Haun A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness, *Consciousness and Cognition*, 79 102877, https://doi.org/10.1016/j.concog.2020.102877.

Tsuchiya, N. (2017). "What is it like to be a bat?"—a pathway to the answer from the integrated information theory. *Philosophy Compass,* 12:e12407, https://doi.org/10.1111/phc3.12407.