Making Sense of Consciousness as Integrated Information: Evolution and Issues of Integrated Information Theory

Kyumin Moon[1], Hongju Pae[2]

[1]*Dept. of Philosophy, Seoul National University*
[2]*Interdisciplinary Program in Cognitive Science, Seoul National University*
dkxnaks@snu.ac.kr[1], hjpae@snu.ac.kr[2]

Abstract

The purpose of this article is to provide an overall critical appraisal of Integrated Information Theory (IIT) of consciousness. We explore how it has evolved and what problems are involved in the theory. IIT is a hypothesis that explains the consciousness in terms of integrated information. It argues that the fundamental properties of experience can be properly analyzed and explained by physical systems' informational properties. Throughout the last decade, there have been many advances in IIT's theoretical structure and mathematical model. Also, like all hypotheses in the field of science of consciousness, IIT has given rise to several controversies and issues. In this context, IIT needs a critical survey. To this end, we first introduce fundamental concepts of IIT and related issues. After that, we discuss major transitions IIT has been through and point out related intra-model issues. Finally, in the last section, some theoretical, extra-model issues involved in IIT's principles are presented. The article concludes by suggesting that, for the sake of future development, IIT should take metacognitive accessibility to experience more seriously.

Keywords: *Integrated Information Theory, the science of consciousness, consciousness, experience, qualia, panpsychism, metacognition*

Journal of Cognitive Science 20-1:1-52, 2019
©2019 Institute for Cognitive Science, Seoul National University

1. Introduction

Integrated Information Theory of consciousness (IIT) is a hypothesis that consciousness can be explained in terms of integrated information. Among other theories, IIT might be one of the most interesting—but also controversial—hypotheses in the field of science of consciousness. IIT is suggested as a principled theoretical framework with explanatory and predictive power. It argues that the fundamental properties of experience can be properly analyzed and explained by physical systems' informational properties. Further, IIT claims that this information-centered, mathematical framework would shed some light on many clinically difficult and ambiguous cases. For this unique approach, IIT has consistently attracted considerable scholarly attention from neuroscientists, information theorists, and even physicists for over a decade. Throughout this period, there have been many advances in IIT's theoretical structure and mathematical model. Furthermore, like all hypotheses in the field of science of consciousness, IIT has also given rise to several controversies (Horgan, 2015; Cerullo, 2015). In this context, there is an urgent need for a critical survey for IIT that would address the following questions: What are the essentials of IIT? What has been changed and what has remained? Moreover, what problems can emerge against them? In the present article, we attempt to provide an overall critical appraisal of IIT.

For this critical review, we will introduce core concepts, major transitions of IIT and related *intra-model* issues that are rooted in the mathematical formulation. Then, several theoretical *extra-model* issues involved in IIT's principles, which are not directly due to the mathematical model, are outlined. Section 2 and 3 cover intra-model, technical issues, while Section 4 deals with extra-model topics. This article concludes by suggesting that, for the sake of future development, IIT should more seriously take metacognitive accessibility to experience.

2. Core Concepts of IIT

Since IIT attempts to explain how conscious experience arises from physical substrates, several explanatory concepts are describing this bottom-up process. While IIT has kept updating its version from 1.0 to 3.0, (Tononi, 2001, 2004, 2008, 2012, Balduzzi & Tononi, 2008, 2009; Oizumi, Albantakis, & Tononi, 2014), those core concepts remain to be fundamentals of the theory throughout all versions. Despite their significant roles in the framework of IIT, the core concepts have not been cashed out. To amend the situation, in what follows, we explain why those concepts are important in IIT and point out some related issues. While our characterization strongly reflects the view of the current version of IIT, providing a summary of IIT 3.0 is not the main concern in this section. Rather, the following descriptions concern several central notions that persist regardless of versions.

2.1. Mechanisms, states, connections, and repertoires

The central focus of IIT is on the physical substrates of experience and their causal structures. IIT analyzes candidate physical substrates of experience in a bottom-up manner; physical elements, which can causally interact with each other, are under consideration. Any set of elements can be considered as a *mechanism.* Furthermore, any set of mechanisms can be thought of as a higher-order mechanism or a *system of mechanisms* (in short, *system*). The system is composed of elements so that the system itself also can be a mechanism or a set of elements. On the other hand, causal structures of physical substrates are analyzed by two central notions of IIT; mechanisms, or systems, can be in a *state*, which corresponds to outputs of their elements. For instance, if three elements—A, B, and C—with the binary output 1 or 0 compose a mechanism, and these element's outputs are respectively 1, 0, and 0, the state of the mechanism ABC is represented as 100 (see Oizumi, Albantakis, & Tononi, 2014, Figure 1A). Further, such a

mechanism in a state can have a *connection*, which corresponds to a set of causal connections among elements of the mechanism (Balduzzi & Tononi, 2009).[1] For example, if causal connections c^1, c^2, c^3, and c^4 are given, there might be a set of connections, such as $\{c^1, c^2\}$, $\{c^1, c^3\}$, $\{c^1, c^2, c^3\}$ or $\{c^1, c^2, c^3, c^4\}$, etc. Any causal relationships could be characterized as a connection, such as synapses between neurons, which could be ideally represented as logic gates with simple computational functions.

From states and connections of the mechanism, one can have *repertoires.* A repertoire is defined as a *probability distribution* to possible states of the mechanism. In IIT, the causal structure of the mechanism must be known *a priori.*[2] When the state and connection of a mechanism are given at time t, one can infer which past or future states of which mechanism–including the mechanism *itself*–could be causes or effects of the given state of the mechanism, and how much probabilities would be distributed to each possible cause or future effect states. Therefore, these probability distributions are probabilistic expressions of how the mechanism's particular state could cause or be caused by a certain mechanism's past or future states. In this sense, the mechanism in the state *specifies* repertoires, or its possible causes and effects.

The notions of mechanisms, states, connections, and repertoires are the very fundamentals in IIT. Without these concepts, calculating information from a mechanism's causal structure is not possible. As explained above,

1 In Balduzzi and Tononi (2009), the term 'submechanism' or 'mechanism' was originally used to refer to sets or subsets of causal connections among elements. However, this use of the term causes a serious confusion, as 'mechanism' is also used in IIT to refer to sets or subsets of elements that causally interact. In order to avoid possible confusions, in the present paper, we use the term 'connection' instead of 'submechanism' or 'mechanism'.

2 This *a priori* known causal structure can be mathematically described by the backward and forward Transition Probability Matrix (TPM) of the set of elements under consideration, which must be known or assessable, e.g. through perturbational methods. While this requirement of *a priori* given causal structure is usually not explicitly presented in literature, it is important and intrinsic to IIT.

repertoires are derived from states and connections of the mechanism. Furthermore, as we will see in Section 2.2, the very concept of information is formally defined by repertoires and related notions. The concepts of mechanisms, states, connections, and repertoires tie causation and information together and enable us to calculate how much information is generated from the causal structure of the mechanism. In part, this is the reason why they survived several updates so far.

These notions also provide IIT with a quite liberal view about possible physical substrates of consciousness. None of these notions tells us about what kind of materials should be considered as a candidate for the physical base of experience. Therefore, when something has its state, connection, and specifies repertories, it can be at least considered as a presence that produces experience. Given that mechanisms or systems in a state are not limited to biological substrates, chemical structures such as silicon chips can be legitimate candidates for the physical base of consciousness. Thus, under the framework of IIT, the question "Is this cellular phone conscious?" is not a category-mistaken question that should be rejected *a priori*. As far as the cellular phone can be considered as a "system of mechanisms in a state," we can at least consider the possibility of its consciousness. In principle, anything that has its states and connections can be a mechanism, and any mechanism can be a possible candidate for a conscious mechanism (Tononi & Koch, 2015).

However, such liberalism comes at a price. While the notions of mechanism, states, connections, *et cetera*, do not limit the *kinds* of physical substrates of experience, they do not limit the *levels* of physical substrates either. Said differently, those basic concepts do not identify in which *spatio-temporal grains* we should find physical substrates of consciousness. Technically, there are elements, mechanisms, systems, state, and connections at each level of the grain; basic particles in microphysical interaction compose quantum mechanisms in a quantum state. Molecules

in chemical bondings constitute chemical mechanisms in a chemical state. Neurons connected with synapses make neuronal mechanisms in a neuronal state. Among these levels, which mechanism should be taken as the origin of consciousness? This applies as same to the macro-levels. For example, in IIT, there appears to be no principled reason not to take China as a single mechanism in a state, composed of causally interacting Chinese people (Schwitzgebel, 2012). Indeed, the problem of finding a proper spatio-temporal grain of consciousness has been admitted by IIT theorists themselves (Tononi, 2008, 2012; Oizumi, Albantakis, & Tononi, 2014). We think that the problem already lies in the center of the basic notions of IIT, leaving theoretical loose ends.

It can be argued that the problem of spatio-temporal grains has already been addressed in the current version of IIT. Applying the exclusion postulate introduced in IIT 3.0, proponents of the theory may argue that the appropriate spatio-temporal grains are ones that have *maximum intrinsic cause-effect power*, which is quantified by the highest value of *integrated conceptual information.* (Tononi, 2012; Tononi, Boly, Massimini, & Koch, 2016) It is nonetheless possible that there are multiple highest values of integrated conceptual information across different spatio-temporal levels. For instance, we might consider a large-scale network involving temporoparietal areas and their connections as well as a mid-scale population of neurons in one of those areas, e.g., the temporoparietal junction. In this case, if both scales produce the same highest values of integrated information at the same time, then which level should be chosen as the level where experience arises? Unless a principled solution is suggested, the problem on pinpointing the proper spatio-temporal grains would still remain.

2.2. Intrinsic and causal information

According to IIT, the amount of information generated by a mechanism is

calculated from repertoires. This calculation is performed by measuring the distance between the unconstrained and constrained repertories. For the past or future state, IIT supposes the unconstrained repertoire as a probabilistic base. Given the system's causal structure, the repertoire is unconstrained in that such uncertainty is not constrained yet by the given state of the mechanism. Using Bayes' theorem, one can infer the constrained repertoire from the given state of the mechanism. It is this distance between unconstrained and constrained repertoires that is defined as information throughout all versions of IIT.[3]

The crucial point here is that those repertoires involved in information should be inferred from mechanisms *within* a considered system.[4] To calculate repertoires specified by the mechanism in the state, one must consider past or future states of mechanisms within the system under consideration. No mechanism outside of the considered system should be taken into account. For example, to calculate the amount of information generated by the mechanism mentioned in Section 2.1—ABC in 100—one should consider mechanisms only from within a considered system. Suppose that the elements A, B, and C is within the system under consideration. Other elements, such as D and E, are outside of the considered system (see Oizumi, Albantakis, & Tononi, 2014, Figure 1A). Then, according to IIT, ABC in 100 cannot specify repertoires of mechanisms such as D, DE, or even AD, AE, ABD, ABE, ABCD, ABCE. It only specifies repertories of mechanisms A, B, C, AB, AC, BC, ABC. Those repertoires would represent possible causes or effects of ABC's

3 For further detail on this calculation, see Section 3.2 and Oizumi, Albantakis and Tononi (2014).

4 The notion of "considered system" or "system under consideration" is explicitly introduced as *candidate set* in IIT 3.0. A candidate set is a *set of elements under consideration*; if certain elements are not included in the candidate set, those elements are considered as external noise, even if they still are part of the whole system. For further details on candidate set, see (Oizumi, Albantakis & Tononi, 2014, Figure 1A).

being in 100 that are in the considered system with their probabilities. In short, mechanisms in a certain system under consideration only can specify repertoires of mechanisms within that system. In this specific sense, in IIT, repertoires specified by the mechanism in a state express *intrinsic causal power* of the mechanism. As repertoires represent the intrinsic causal power of the mechanism, information in IIT is essentially *intrinsic* and *causal*. Information generated by the mechanism is measured as the distance between repertoires. Of note, these repertoires involve nothing external to the system. They solely depend on possible causes or effects within the system. Therefore, information is intrinsic to the system in that it does not require anything external to the system. Also, information has nothing to do with input/output signals that can be detected only by the external observer. Rather, it is about the causes and effects that can be detected only from *the system's intrinsic perspective* (Tononi, 2008, 2012; Oizumi, Albantakis, & Tononi, 2014). Moreover, given that repertoires specify possible causes or effects and their probabilities, information produced by the mechanism is causal. At this point, IIT repeatedly emphasizes the notion of information as *"differences that make a difference"* (Bateson, 1972). In IIT, for instance, the mechanism in a state specifies which past states of a certain mechanism ("differences") would likely to cause the mechanism's being in that state ("a difference"). This further implies that only something that can be *selectively* caused or cause can produce information. This intrinsic and causal notion of information is the hallmark of IIT, which distinguishes IIT from other information theories: anything informative has intrinsic causal power, and anything intrinsically causal has information. This intrinsic and causal nature of information is directly inherited by the most central concept in IIT, integrated information. Although integrated information is defined in a sophisticated manner, in so far as it is information, it also should be intrinsic and causal. The intrinsic and causal information is fundamental to IIT in that it determines what kind of information the theory deals with.

Although intrinsic and causal information constitutes one of the unique aspects of IIT, it also raises some problems concerning the *function of consciousness*. Simply put, intrinsic and causal information does not involve anything outside of the system. By definition, integrated information has nothing to do with *causal inputs/outputs* of the system, too. This intrinsicness renders integrated information irrelevant to the functions of the system. As we will see in Section 4.2., IIT theoretically designs functional zombie systems, which share all the input-output relations with systems with highly integrated information. This suggests that integrated information is nearly irrelevant to the functions of the system; integrating information has no necessary bearing on the system's functioning (Schwitzgebel, 2014). In the sections below, we will see that IIT identifies integrated information and consciousness. If so, integrated information's functional irrelevancy would be directly transferred to consciousness. For instance, IIT implies that, at least in principle, there can be perfectly functional equivalents of *us* that are unconscious. It is at least theoretically possible that we do whatever we are doing without consciousness. Such state makes the 'use' of consciousness as mysterious. Moreover, the *adaptive benefits* of having experience also become doubtable; whatever adaptive function experience provides, there is always a possible scenario that the function might have been evolved without experience. This risk of the functional irrelevancy of experience has been already rooted in the intrinsic nature of information in IIT.

2.3. Integrated information and complex

The notion of *integration* first stems from the phenomenological aspects of experience: "[p]henomenologically, every experience is an integrated whole, one that means what it means by being one, and which is experienced form a single point of view" (Tononi, 2012, p. 295). To be a physical underpinning of such integrated, unified experience, what should

a mechanism be? Here, IIT suggests one of its thought experiments: let's compare a highly informative, but unconscious mechanism and a conscious mechanism. For example, what is the difference between a conscious brain and an unconscious digital camera that consists of thousands of photodiodes? According to the IIT, the most significant difference is that while the former is causally integrated, the latter is not (Tononi, 2012). Causal interactions within the brain are so highly integrated that, once they are fragmented, the whole brain's performance might break down. This thought experiment on the camera model suggests that producing information is not sufficient for a mechanism to generate consciousness. Even if the mechanism is equipped with complicated connections and distinguishes vast repertoires, if its elements do not specify a maximum of integrated information, the mechanism cannot give rise to experience.

As a mechanism with a causal structure produces intrinsic information, one with integrated causal structure generates integrated information. The integrated information is integrated in the sense that, as a whole, the mechanism generates more information than the sum of its parts. Said differently, it is information produced only from the mechanism as a whole. By definition, the integrated information of the system is irreducible to its parts. Therefore, according to IIT, the amount of integrated information generated by the mechanism is calculated by partitioning the system by disconnecting the connections between the mechanisms. That is, if the information disappears by partitioning, it would be the information generated by the mechanism as a whole, not by individual parts. The informational difference between the mechanism as a whole and the system's partitions' mechanism is defined as integrated information.[5] Nonetheless, considering that there are many possible ways of how the mechanism is partitioned, it becomes crucial to decide which partition

5 This idea of integration might be closely related to the notion of synergy information proposed by Virgil and Koch (2014).

should be used in calculating integrated information. IIT chooses the partition which causes the least loss of information, which is called a *minimum information partition* (MIP). Finally, depending on the level of calculation, the calculated values of integrated information are represented as Φ or φ.

Based on the integrated information, a complex is defined: roughly put, parts of the system producing integrated information can be considered as complexes.[6] As we will see in Section 2.4, IIT posits the identity between consciousness and integrated information; complexes in the system directly contribute to consciousness by integrating information. Technically, only complexes should be regarded as physical substrates of conscious experience, and they deserve to be called a 'locus' of consciousness. Despite a significant change concerning whether the overlapping or inclusion among complexes is possible, IIT maintains that a system can be condensed into multiple complexes. Finding such complexes in the system is the main focus of IIT on defining the local and temporal origin of consciousness.

It should be emphasized that the notions of integrated information and complex provide possible explanations for some fundamental properties

6 Technically, how complexes are defined depends on which version of IIT is taken. A complex is defined as a set of elements which produces a local maximum of integrated conceptual information on a system level, which quantified by Φ^{max} in IIT 3.0. While this is true, strictly speaking, notions such as integrated conceptual information, local maxima, and mechanism-system distinction were explicitly introduced since IIT 3.0. One cannot find any of these before IIT 3.0. In the IIT 2.0, complexes are defined differently, as sets of elements that produce integrated information. As we have noted at the beginning of section 2, our purpose is not briefly presenting the current version of IIT. (If it was, we would not cite any of the literature based on IIT 2.0 framework, including (Tononi, 2008) or (Balduzzi & Tononi, 2008, 2009)). The focus is on the core concepts that play essential roles *throughout* all versions of IIT. Thus, until the end of section 2, we temporarily choose to ignore conceptual differences among various versions and use some terms very loosely. In section 2, for instance, "integrated information" covers integrated information in IIT 2.0 as well as integrated conceptual information and maximally integrated conceptual information in IIT 3.0. Accordingly, "Φ" can refer both Φ and $\Phi^{max.}$

of experience. In calculating integrated information, nothing outside of the complex matters. Whether how much the information is generated by the complex, or how much the information is lost by MIP, is purely intrinsic to the complex. This intrinsicness of integrated information accounts for why experience is essentially intrinsic. In other words, the experience is integrated information, and integrated information is intrinsic. Therefore, the experience is intrinsic. This characteristic of experience can also be noted as *privacy:* "Since integrated information is generated within a complex and not outside its boundaries, the experience is necessarily private and related to a single point of view or perspective" (Tononi, 2008, p. 221). Appealing to the central concepts such as integrated information and complex, IIT appears to open up the prospects of making sense of essential features of experience.

While integrated information is undeniably the key concept in IIT, it also creates a problem which renders the application of IIT to real systems practically intractable. As specified above, calculating integrated information involves finding MIP, which requires creating all possible partitions and measuring all the informational difference between the non-partitioned and the partitioned. With the growth of the number of elements organizing the system, it becomes obvious that the amount of computation will dramatically increase. Consequently, one faces a serious *combinatorial explosion* in finding MIP.[7] Due to this computational burden, applying IIT to neural substrates or artificial robots is currently infeasible. At the current stage of the theory, since direct empirical data supporting IIT are unavailable, researchers have tried to find efficient algorithms for

7 Here, the levels of MIP need to be distinguished: the MIP on the level of small phi and the MIP on the level of conceptual information big Phi. It is of course true that these are two different forms of partitions that are respectively applied to the levels of mechanism and system. However, as clarified in footnote 6, such notions are restricted to the current version. They cannot be applied regardless of versions.

choosing MIP (Kitazono, Kanai, & Oizumi, 2018; Hidaka & Oizumi, 2018), or to develop approximations or proxy measures of Φ.[8] The absence of experimental validity is a decisive disadvantage for IIT to become a solid theory claim to the science of consciousness.

2.4. Identity between consciousness and integrated information

As mentioned in Section 2.3, IIT identifies consciousness with integrated information in the first place. Particularly in IIT, levels of consciousness are identified with *quantities* of integrated information, while *qualities* of consciousness are identified with *informational structures* derived from integrated information. From these identifications, IIT attempts to account for both how conscious a system is and how it feels.

According to IIT, a level of consciousness is nothing but an amount of integrated information. Therefore, one can know "how conscious the system is" by calculating the amount of integrated information produced by that system.[9] Consciousness is not all-or-nothing. Rather, as shown by the

8 A large variety of modified Φ has been proposed as an estimation for Φ. $\tilde{\Phi}_E$ is modulated by the Markovian discrete system and can be applied to continuous time series data (Barret & Seth, 2011), and Φ* is modulated by substituting the notion of decoding perspective of information that facilitates the overall computation procedure (Oizumi, Amari, Yanagawa, Fujii and Tsuchiya, 2016). However, they do not contain the main theoretical updates of the IIT, such as cause-effect information and the distinction between φ and Φ. For IIT 3.0, Marshall, Gomez-Raminez, and Tononi (2016) have proposed State Differentiation (SD) as a proxy measure of Φ, which is much easier to draw out from experimental data than the original Φ. Still, it leaves the degree of integration being not properly measured. It is also insufficient to assume that SD functions are a complete form of measure in that it is applied to cellular animats; it is still analyzing the toy problem of the causal system. Recently, Tegmark (2016) has proposed several kinds of modified Φ through various definitions of informational distance and by normalizing integrated information using diverse techniques.

9 Rich integration of neuronal connection is widely known as a major feature of the cerebral cortex, based on the anatomical structure of the brain. The fact that the brain cortex is constructed as a complicated neuronal network can explain how consciousness arises from such anatomical structure and why there exists such direct correlation between Φ and consciousness.

experience of falling asleep or that of anesthesia, consciousness is a matter of the degree. IIT claims that a level of consciousness can be quantitatively measured by a value of integrated information, which is referred to as Φ. Since this 'quantifying consciousness' has drawn considerable scholarly attention, many IIT studies thus far have been dedicated to finding correlations between levels of consciousness and corresponding Φ values. Some of the results from analyzing EEG data and computer simulations suggest that Φ can be a reliable measure of consciousness (Tononi, 2008). Indeed, the idea of the possibility to measure consciousness quantitatively alludes to the science of conscious experience. It is the identification of levels of consciousness and Φ values that makes this idea possible.

On the other hand, IIT claims that quality of experience is just an informational structure assessed by integrated information. This informational structure can be represented as a geometrical shape in a multidimensional space that "completely and univocally specifies the quality of experience" (Tononi, 2008, p. 224). If so, how the system feels can be known by deriving what shape is represented by the integrated information it generates. Each version of IIT provides sophisticated procedures for illustrating shapes like polytopes on multidimensional space from given integrated information. These shapes specify informational relationships generated by complexes. It seems to be obvious that, if it is successful, this geometrical representation provides useful tools for analyzing qualities of experience. Qualities of experience have several fundamental aspects to be explained, such as similarities and differences, richness, heterogeneities, as well as compositional structures. Once qualities of experience are identified with shapes assessed from integrated information, those fundamental aspects can be explained by analyzing geometrical characteristics of shapes in the multidimensional space. This explanatory potential of geometrical approach might be the most distinctive part of IIT: "what it is like to experience" could be explained by the "geometry of integrated information"

(Balduzzi & Tononi, 2009).

The identity between consciousness and integrated information has further implications. First, in clinical contexts, quantifying consciousness by Φ value might play a significant role in treating pathological cases of patients in coma or vegetative state. As the locked-in syndrome case suggests, how to judge whether or not one is conscious has been an extremely controversial issue. However, if Φ is indeed a level of consciousness, we have a simple answer: a patient is conscious only when his or her brain generates non-zero Φ. This answer immediately leads to a liberal approach to the consciousness of non-humans. For animal consciousness, animals can be conscious if they generate integrated information at all. The same applies to artificial consciousness, as there is no reason to *a priori* exclude the possibility that artificial intelligence can be conscious. The only thing that matters upon consciousness is whether or not the candidate system produces non-zero Φ (Tononi & Koch, 2015). Second, the idea of qualities of experience as geometrical shapes appears to entail that experience is *substrate-independent*. According to IIT's geometrical approach, systems with different states and connections can produce the same informational structures (Balduzzi & Tononi, 2008, 2009). From the assumption that qualities of experience are nothing but informational structure geometrically depicted from integrated information, it follows that qualities of experience and their physical substrates can come apart. This substrate-independence might account for, at least partially, why consciousness seems non-physical (Tegmark, 2017). In this respect, many theoretically and practically promising predictions and explanations come from the identification of consciousness and integrated information.

However, the notion of consciousness as integrated information also raises some perplexing issues. Since the theory identifies Φ with a level of consciousness, IIT must ascribe experience to seemingly unconscious systems. Surprisingly, according to IIT, a photodiode with the binary states,

on or off, is minimally conscious, because it produces non-zero Φ.[10] What is more, a lattice structure composed of a single kind of logic gates can be highly conscious, even much more conscious than the human brains. In this way, IIT predicts that a functional zombie is empirically possible in principle (Oizumi, Albantakis, & Tononi, 2014); even if two systems are functionally equivalent, there can be a situation when one generates Φ, but the other does not (see Oizumi, Albantakis & Tononi, 2014, Figure 21). The theoretical development of IIT is not without a sense of irony: its core concepts promise ample explanatory and predictive potentials but also brings counterintuitive and problematic defects to the theory. In the remaining chapters, these intriguing issues will be analyzed further in detail.

3. Major Transitions in IIT

While its core ideas are more or less preserved, IIT has undergone several significant revisions from its prototype to the very latest version. These revisions made the framework of IIT more theoretically and technically articulate. Not only its theoretical structure but also the details of its mathematical model have been changed. Therefore, understanding how IIT has acquired its current form requires a deeper analysis. Axioms, postulates, and the notion of purviews have been introduced, the concept of information has been revised, the distance metric for repertoires has been changed, and levels of information integration have been divided. In what follows, we explain what these changes are about and how they affect IIT. While until now we have set aside the revision of IIT 3.0 in the review, as

10 Again, the reason why a photodiode should be treated as minimally conscious can be different according to versions of IIT. In IIT 2.0, the photodiode is minimally conscious, since it produces one bit of integrated information. However, in IIT 3.0, it is so because it generates non-zero integrated conceptual information. Although this distinction is important, for the purpose of section 2, we do not deal with differences among versions. See footnote 6.

this section is about the *transitions* between the previous and the current version, we will deal with IIT 3.0 from now on.

3.1. From thought experiments to systematic formulation: Phenomenological axioms and ontological postulates

As occasionally mentioned in the literature, the question of what consciousness is and what a physical system should be to generate consciousness has never been explicitly answered until IIT has developed into its latest version. The fundamental properties of experience were taken for granted by appealing to phenomenology, and the required properties for physical systems to produce experience were motivated or suggested merely by thought experiments. The photodiode and camera thought experiments were introduced from the early version of IIT (Tononi, 2004, 2008), and the Internet thought experiment was added during the updates to IIT 3.0 (Tononi, 2012). Based on the idea that conscious experience is specific in a particular way, the photodiode thought experiment suggests that physical systems must specify their possible causes and effects that generate consciousness. The digital camera thought experiment suggests that physical systems must be causally integrated since it appears that experience is unified and integrated. By contrasting information of the Internet and that of experience, the Internet thought experiment speculates that information produced by physical systems must be maximally integrated. While these thought experiments are interesting in themselves and might be helpful to understand the motivations behind the theory, they never clearly argue for or even identify the fundamental features of the essential properties of experience.

In IIT 3.0, the situation has changed. Now, the fundamental properties of consciousness and requirement for physical systems are explicated and posited in the very beginning of the theoretical formulation. First and foremost, *phenomenological axioms* are introduced; these axioms

are phenomenological in the sense that they all concern the fundamental properties of experience. Each of the five axioms corresponds to each of the essential properties of experience: the existence axiom states that consciousness exists; the *composition* axiom says that it is compositional; the *information* axiom states that it is informative; the *integration* axiom claims that it is integrated; the *exclusion* axiom says that one consciousness excludes another consciousness (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). These properties in the axioms are supposed to be fundamental, as any experience must have them.

Next, corresponding to the phenomenological axioms, *ontological postulates* are posited: these postulates are ontological in that they prescribe what mechanisms should generate consciousness. There are five postulates which lay parallel to each axiom. The *existence* postulate says mechanisms in a state must exist. The *composition* postulate says that mechanisms must be structured. The *information* postulate claims that mechanisms must produce information by specifying selective possible causes and effects within the system. The integration postulate states that mechanisms must integrate information. Finally, the *exclusion* postulate says that mechanisms must generate only the maximally integrated information (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014; Tononi & Koch, 2015). As in the phenomenological axioms, properties mentioned in ontological postulates are essential to *every* physical mechanism to generate consciousness. Moreover, the latter three postulates—information, integration, and exclusion—are applied to the two different levels of calculation; the *mechanisms* and the systems *of mechanisms*. Contents of postulates vary through systems, depending on which level they are postulated.[11]

In virtue of these axioms and postulates, IIT becomes a top-down and theory-driven approach, rather than as a bottom-up and experiment-driven

11 For further details on ontological postulates, see Oizumi, Albantakis, & Tononi (2014).

approach to consciousness; axioms and postulates come first and later comes the mathematical model. Empirical experiments can be designed and conducted only under the models and theories. Consequently, by declaring its axioms and postulates, IIT can clarify both its theoretical framework on modeling the consciousness and the experiments.

However, while clarification is one thing, justification is another thing. The introduction of the axioms and postulates raises many questions. First, on what ground must phenomenological axioms be accepted? That is, why should those axioms be considered axiomatic? The axioms themselves appear to be based on phenomenological intuition or introspection. The axioms are "assumed to be self-evident from the intrinsic perspective of a conscious entity" (Oizumi, Albantakis, & Tononi, 2014, Supplementary 1, p. 1). However, what if such intuitions or introspections are wrong? Moreover, how can each ontological postulate follow from each phenomenological axiom? Though the postulates are consistent with the axioms, there seems to be an unbridged gap between them. For instance, it is not clear how we can draw the information postulate; it is not clear if the mechanisms should specify selective causes and effects within the system from the information axiom, which states that consciousness is informative. Therefore, the rationales for positing phenomenological axioms and ontological postulates remain controversial. Recently, Bayne (2018) addresses precisely the axiomatic foundations of IIT. Bayne argues that some of the phenomenological axioms are not self-evident, and others seem to be self-evident but fail to practically or theoretically constrain the theory of consciousness at all. He suggests that IIT would be on firmer ground if it adopts what he calls 'natural kind approach' (Bayne, 2018). While the verdict may still be out, it appears that the axiomatic approach and seemingly following postulates are not so secure as they seem.

3.2. From effective information to cause-effect information: New information and its metric

The revision of the notion of information might be one of the most significant developments during the updates of IIT. In the early versions of the theory, information generated by a system was defined as *effective information (ei)* (Tononi, 2008). There were two kinds of repertoires: a potential repertoire, which is a probability distribution of the past states when no current state of the system is known, and an actual repertoire, which is a probability distribution of the past states when a particular current state of the system is known. One can measure the distance between the potential and the actual repertoires by applying *Kullback-Leibler Divergence (KLD)*, and such distance could be thought of a sort of *relative entropy.* This distance or relative entropy directly equals the effective information. In IIT 3.0, however, a new form of information is introduced: *cause-effect information (cei)* (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). The cause-effect information differs from effective information in many important aspects.

First, unlike ei, cei involves the system's past *and* future (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). In calculating ei, potential and actual repertoires are only of the past states of the system. In calculating cei, however, repertoires concern *both* the past *and* future states of the system. These repertoires can be thought of as probabilistic expressions of how the past states of the system would cause the current state of the mechanism and how it would cause the future states of the system. On the one hand, there are *unconstrained past repertoire* and *cause repertoire.* The former is a probability distribution of the past states of a certain mechanism of the system when the current state of the given mechanism is not known. It always produces maximum entropy of past states. The latter is defined as a probability distribution of the past states of a certain mechanism of the system when the current state of the given mechanism is known. On the

other hand, there are *unconstrained future repertoire* and *effect repertoire*. The former is a probability distribution of the future states of a certain mechanism when the current state of the given mechanism is perturbed in every possible way. The latter repertoire is a probability distribution of the future states of a certain mechanism of the system when the current state of the given mechanism is known. The current state of the given mechanism specifies cause and effect repertoires of a certain mechanism of the system. It should be noted that unconstrained past and future repertoires and cause and effect repertoires are calculated *independently* of each other. Therefore, repertoires should be calculated *twice* in calculating cei; therefore, while calculating ei requires only two repertoires, four repertoires are required to calculate cei.

Second, while ei concerns only the states of the given system itself, cei can involve the past/future states of mechanisms *other than* the given mechanism (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). For ei, only the past states of the same system should be taken into account, as potential and cause repertoires are defined as probability distributions of the past states of the system itself and nothing else. However, in calculating cei produced by the mechanism, not only the past/future states of the mechanism itself but also those of other mechanisms of the system can be considered. Said differently, the repertoires required to calculate cei of the given mechanism are not restricted to the same mechanism. For example, if the whole system is composed of elements A, B, and C with binary outputs 1 and 0, and the selected mechanism is A, A's current state 1 can specify the cause repertoire of the past states of any mechanism, including B, BC, AC, ABC, or even A itself. Similarly, it can also specify the past repertoire of the future state of any mechanism (see Oizumi, Albantakis, & Tononi, 2014, Figure 4). Thus, to calculate cei generated by A in 1, one must first decide which mechanism to be *paired* with A. In principle, any mechanism of the system, which could be represented as the power set of the total elements

of the system, can be paired with AB. In IIT 3.0, this idea of pairing is introduced as *purview*. When A in 1 is paired with ABC and the cause repertoire is calculated, the purview of A is represented as A^c/ABC^p. If it is paired with ABC and the effect repertoire is calculated, the purview of A is represented as A^c/ABC^f (see also Oizumi, Albantakis, & Tononi, 2014, Figure 4). Once the purview is fixed, other elements outside the purview remain unconstrained and do not affect cause and effect repertoires. However, calculating unconstrained repertoire does not require a specific purview, because there is no difference in unconstrained repertoires among different current purviews. By discriminating the mechanism's purview, the causal analysis could be extended to every mechanism of the system.

Third, whereas ei is measured by KLD, cei is measured by a different metric. As explained above, to calculate ei, we must measure the distance between the potential and the actual repertoires. In the earlier versions, it was KLD which was used to measure the distance. KLD is the most intuitive index for measuring the reduction of entropy, which directly relates to the quantity of information generated in terms of relative entropy. Since entropy and information were regarded as symmetrical, KLD was chosen for the scale of distance during the early versions of IIT. However, technically, KLD should not be considered as a proper metric, since it is not symmetric, does not obey triangular inequality, and is unbounded. Also, non-compensated KLD measures only the reduction of uncertainty and does not account for the difference between states, which appears to be crucial in calculating information. For these reasons, another measure should be introduced as a new scale (Oizumi, Albantakis, & Tononi, 2014, Supplementary 2).

Therefore, from IIT 3.0, *Earth Mover's* Distance *(EMD)* is used to measure the distance between repertoires. This is also known as Wasserstein distance, which is the distance function defined by the minimum cost of redistributing the "dirt piles" to the location elsewhere (Oizumi, Albantakis,

& Tononi, 2014, Supplementary 2). Given that distributed probabilities can be thought of as "dirt piles," one can think of a distance between two repertoires as the minimum cost of distributing "dirt piles." In IIT, there are two kinds of EMD. First, a general EMD is applied in calculating cause-effect and integrated information on the level of mechanisms. Second, an extended EMD is used to calculate that on the level of the systems of mechanisms. Nevertheless, in both cases, the point of using EMD remains the same. By using EMD, not only the reduction of entropy but also the difference between states is taken into account in calculating information. From IIT 3.0, the quantitative value of information is not represented by a bit, since the unit of the distance measured by EMD is not a bit. In sum, EMD appears to be a more appropriate metric for IIT than KLD. Based on EMD, the distance between unconstrained past repertoire and cause repertoires is defined as the *cause information (ci)*. The cause information illustrates possible causes of the mechanism's current state when a purview of the mechanism is fixed. Similarly, the distance between unconstrained future repertoire and effect repertoire is defined as the *effect information (ei)*. This implies that effect information signifies possible effects of the mechanism's current state when a purview of the mechanism is fixed. In sum, effect information can be calculated from the distance between those repertoires and is quantified by EMD as same as cause repertoire.[12]

Finally, there is an informational principle that should be applied to cause-effect information. In the earlier versions of IIT, measuring the distance between repertoires was all that mattered. The measured distance was the amount of ei of the system. In IIT 3.0, however, there is more than just measuring the distance. As explained so far, two pairs of repertoires lead to two kinds of information: cause and effect information.

12 After IIT 3.0, a large variety of distance functions, such as Hilbert-space distance and Shannon-Jensen distance, have been newly proposed as metrics of informational difference (Tegmark, 2016). It would be important to consider the characteristics of each measure in order to broaden the explanatory power of IIT.

Then, which information should be accounted for as the mechanism's information? At this point, the *Information Bottleneck Principle (IBP)* is introduced (Oizumi, Albantakis, & Tononi, 2014). IBP forces one to choose the *minimum of* cause and effect information. The motivation behind IBP comes from the intrinsic and causal notion of information: since information in IIT is supposed to be intrinsic to the system, the information that can be detected only by the external observer must be excluded. Suppose if the mechanism in a state only generates cause information, but no effect information. This implies that the mechanism being in such a state does not make any difference to the system. In that case, although the mechanism still belongs inside the system, it does not give any causal interaction among the system's other mechanisms. Hence, such cause information produced purely by the mechanism cannot be detected from the intrinsic perspective of the system. The same holds when the mechanism in the state produces only effect information, but no cause information. This observation enforces IBP so that the smaller one between cause and effect information is taken as cei.

Since the concept of information is at the heart of the theory, the transition from ei to cei articulates the framework of IIT in many important aspects. By taking into account both the past and the future, the notion of information becomes more causal; it involves causes *and* effects. The application of IBP makes it more intrinsic. Specifically, as we will see in Sections 3.3-3.4, when it comes to integrated information on the level of mechanisms, considering all possible purviews of the given mechanism plays a crucial role in enforcing the exclusion postulate. This involves the central notions of IIT 3.0, including concepts, conceptual structure, and other related ideas.

However, the technical complexity is the other side of the theoretical articulation. As mentioned above, the computational burden is doubled, since repertoires and information must be calculated twice. The

multidimensional space for geometrical representation of concepts is also doubled (see Oizumi, Albantakis, & Tononi, 2014, Figure 15). Moreover, calculating integrated information of mechanisms becomes more computationally complicated, because it should concern possible purviews of the mechanism. To calculate integrated information of the mechanism, MIP should be found in each possible purview (see Oizumi, Albantakis, & Tononi, 2014, Figure 8). In this sense, it appears to be clear that the introduction of purview worsens the combinatorial explosion. In what follows, all these technical issues will be analyzed in further detail.

3.3. From one phi to two phis: the distinction between φ and Φ

Before IIT 3.0, there was only one kind of integrated information; all integrated information calculated from the system was noted as Φ. Cause-effect repertoires were inferred from the mechanism, and it was all that mattered in calculating integrated information. However, from IIT 3.0, the distinction between the level of mechanism and that of systems of mechanisms has been introduced. According to this distinction of levels, a distinction between kinds of integrated information has been made (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). On the one hand, there is *integrated information* generated from mechanisms, which is indicated as φ (small phi); on the other hand, there is *integrated conceptual information* produced by systems of mechanisms, which is represented as Φ (large phi). φ and Φ differ from each other both in their concepts and calculations.

Integrated information φ succeeds the motivation *"more than the sum of its parts"* from older versions of IIT and is still analyzed on mechanisms. According to IIT, if there is a difference between the sum of the cause-effect information created by the partition of mechanism and the cause-effect information generated by the unpartitioned mechanism, and this difference directly refers to the information which mechanism forms as a whole entity. Any possible subset of a mechanism which can make a

difference on repertoire could be a candidate for partition.[13] On the level of the mechanisms, φ can be measured by making a partition on a given purview; for example, the purview of ABC in 100 is defined over the past mechanisms in a state. As briefly explained in section 2.3, among all possible partitions, MIP is selected for the calculation of integrated cause information, φ_{cause}. The purview of ABC also can be defined over the future mechanisms in a state. Applying the same procedure, integrated effect information, φ_{effect}, can be calculated. By IBP, one can have integrated information $\varphi_{cause\text{-}effect}$. In this way, φ can be calculated respectively from every single purview available on a certain mechanism (see Oizumi, Albantakis, & Tononi, 2014, Figure 8).[14]

However, since there can be many possible past and future purviews on a mechanism, one mechanism in a state can have multiple possible φs. Here, one of the ontological postulates comes in: *the exclusion postulate* states that, in order to contribute to experience, a mechanism must have only one set of possible causes and effects which is maximally irreducible, while all other sets should be excluded (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). It means that only the cause-effect repertoire of the mechanism that provides the maximum value of φ, φ^{max}, should be taken. As φ is defined as the minimum of φ_{cause} and φ_{effect}, in order to find φ^{max}, one must find the maximally irreducible cause repertoire that yields $\varphi_{cause}{}^{max}$ and the maximally irreducible effect repertoire that provides $\varphi_{effect}{}^{max}$ first. Then, the minimum of $\varphi_{cause}{}^{max}$ and $\varphi_{effect}{}^{max}$ would be φ^{max}. The maximally irreducible cause repertoire is called a *core cause*; the maximally irreducible effect repertoire *core effect.* The pair of core cause and effect is called *Maximally*

13 It is interesting that one of the variable subsets at a certain time might be empty as a result of a particular partition. Furthermore, partitioning can be thought of a method of making certain mechanisms causally inactive. This process is called 'virtualizing the element' or 'injecting noise to the mechanism'. For more detailed analysis, see (Krohn & Ostwald, 2017).

14 For the details of these computational steps, see (Oizumi, Albantakis, & Tononi, 2014, Figure 6).

Irreducible Cause and Effect repertoire (MICE). MICE or the mechanism which specifies MICE is called a *core concept*, or just *concept*.[15] In short, by the exclusion postulate, the highest value of φ should be chosen among all possible φs produced by the mechanism and is defined as φ^{max}. Here, the mechanism that produces φ^{max} is should be regarded as the concept.

Given concepts, one can calculate the amount of integrated conceptual information at the level of systems of mechanisms. Concepts can be illustrated as points in a multidimensional space called *concept space*, and these points would make 'constellations' in the coordinate space. In IIT 3.0, the constellation of concepts is defined as a *conceptual structure* (Oizumi, Albantakis, & Tononi, 2014). As each mechanism specifies its MICE, the system of mechanisms specifies its conceptual structure in the concept space. From this conceptual structure, one can calculate *Conceptual Information (CI)* generated by the system of mechanisms. As CI corresponds to the cause-effect information, it is quantified similarly; as there must be unconstrained past and future repertoires for calculating cause-effect information, there must be the *"null" concepts* for calculating CI. "Null" concepts are unconstrained past and future repertoires in which the state of the system of the mechanism is undecided.[16] By applying the

15 In IIT literature, the use of the term *concept* seems inconsistent: on one hand, 'concept' seems to refer MICE, a maximally irreducible cause-effect repertoire. In (Oizumi, Albantakis, & Tononi, 2014), it is said: "the notion of a *concept*: the maximally irreducible cause-effect repertoire of a mechanism" (ibid., p3). On the other hand, it is also used to indicate mechanism which specifies the MICE: "If the MICE exists, the mechanism constitutes a *concept*." (ibid., p3) "A mechanism that specifies a *maximally irreducible cause and effect(MICE)* constitutes a *concept*" (ibid., p9). Also see (ibid., p5, Table 1). The term is even described as denoting both. "Concept: A set of elements within a system and the maximally irreducible cause-effect repertoire it specifies, with its associated value of integrated information(φ^{max})" (ibid., p5, Box 1) To avoid possible confusions, we choose the second use. In this article, the term concept will always refer to the mechanism specifying MICE.

16 The "null" concepts are named so because they specify unconstrained past and future repertoires if considered as mechanisms; in other words, it is the concept

extended version of EMD, it can be quantified how much CI is produced by the system of mechanisms.[17]

The calculation of integrated conceptual information Φ is also analogous to that of φ. Once the purview of the system of mechanisms is given, MIP can be found by partitioning[18] the purview. By measuring the difference of CI between the unpartitioned and the partitioned, MIP can be calculated how much the system of mechanisms generates integrated conceptual information. Again, as for φ, there can be many possible Φs as all possible unidirectional partitions of the set of elements should be considered. Here, the exclusion postulate comes in again: it enforces only one complex among all other overlapping systems of mechanisms to contribute to consciousness. Thus, the one that generates the maximum of Φ should be chosen as Φ^{max}. Finally, the conceptual structure that gives rise to Φ^{max}

that specifies nothing. Although it is perceived as only a superficial notion on designating unconstrained repertoire, however, it also can be illustrated in conceptual space along with other concepts (see Oizumi, Albantakis, & Tononi, 2014, Figure 11). By this way, "null" concept refers to the concept specifying its current purview of mechanism as an empty set.

17 Extended EMD differs from original EMD by its methods on calculation. The distance between cause-effect repertoire and unconstrained cause-effect repertoire is measured by EMD, and each EMD of cause and effect distributions are added up, then it is multiplied by φ^{max} of the concept, which functions as the weight of each concept. In short, extended EMD is used on the level of systems of mechanisms by multiplying φ^{max} as each distribution's weight. Even if the details on the calculation vary, the fundamentals on calculating the distance, which is redistributing the probability distribution, do not change. For further detail on applying extended EMD, see (Oizumi, Albantakis, & Tononi 2014, Supplementary 2).

18 The partitioning in the level of systems of mechanisms must be *unidirectional* (Oizumi, Albantakis, & Tononi, 2014). Unidirectional partitioning is done by virtualizing elements; when a mechanism is injected with noise, information disappears, as the mechanism gets considered as external noise and loses its intrinsic causal power. Unidirectional partitioning could be thought of as injecting noise between subsets for only to a certain direction of the connection. Thus, partitioning the direction of connection between subsets on the system level is analogous to virtualizing the elements on the mechanism level. For further detail on virtualizing the elements, see also Krohn and Ostwald (2017).

is defined as *Maximally Integrated Conceptual Structure (MICS)*. The system of mechanisms that produces Φ^{max} and so specifies MICS is defined as *complex*.[19] In IIT 3.0, such MICS generated by the complex is directly identified as the subjective experience.

By introducing the distinction between φ and Φ, now it has become much logical to explain the generation of consciousness using integrated information in further detail. In IIT 3.0, MICE is called *quale sensu stricto*, which means quale in the narrow sense. Since this sort of quale includes 'redness of red' or 'painfulness of pain,' it can be considered to be quale in the philosophical debates. On the other hand, MICS is called *quale sensu lato*, which means quale in a wide sense (Oizumi, Albantakis, & Tononi, 2014). As mentioned above, IIT equates experience with MICS. As mechanisms in the complex maximally integrate cause-effect information, concepts are generated, and we have qualia. As the complex maximally integrates CI, MICS is produced, and we have an experience. Based on the distinction and the exclusion postulate, the notions of concepts and MICS

19 As a result, there comes an important change in defining the complex since IIT 3.0. In IIT 3.0, due to the exclusion postulate, complexes cannot be nested or overlap at all. In the earlier versions, however, since the exclusion axiom/postulate was not introduced yet, complexes could partially or wholly overlap. Meanwhile, there can be multiple complexes in one system. IIT predicts that one system can be condensed into *several* complexes. The complex that has Φ^{max} is called *major complex*, and the complex which does not overlap, but has Φ smaller than Φ^{max} is called *minor complex*. Since they have their own Φ^{max}, they are considered as an individual complex. Minor complex can be thought of as a local maximum which implies 'locally condensed minimal consciousness' (see Oizumi, Albantakis, & Tononi, 2014, Figure 16). There should be a single major complex in general situations, but there could be multiple major complexes according to circumstances. For example, split brain syndrome or dissociative disorders could be explained as clinical examples of the main complex being split into two or more. At the same time, minor complexes could be thought of as *preconsciousness*; the constituent of consciousness which can contribute to the reaction of extrinsic inputs. Continuous flash suppression could also be explained through the function of the minor complex (Oizumi, Albantakis, & Tononi, 2014).

can be defined. These central notions of IIT 3.0 enable one to explain how experience arises from its physical substrates in a bottom-up manner. All these articulated explanations essentially start from the distinction between φ and Φ.

However, this distinction between φ and Φ also raised several problems for IIT. First and foremost, computing Φ and the application to real systems became computationally intractable (Oizumi, Albantakis, & Tononi, 2014). Almost every aspect of calculation doubled: MIP had to be found twice; once at the level of mechanisms and twice at the level of systems of mechanisms. Owing to the distinction between φ and Φ, it appears that the combinatorial explosion in IIT worsened. In turn, such computational infeasibility rendered the empirical prospect of IIT more unpromising.[20] At the cost of an articulated bottom-up explanation of experience, the theory had to face serious practical problems in retaining empirical validity.

Another problem emerges from the exclusion postulate. As explained above, the exclusion postulate enforces that only the mechanisms which give rise to φ^{max} or the systems of mechanisms which provide Φ^{max} must be taken as the concept or complex. However, if several different MICEs yield the same φ^{max}s, then which repertoire or conceptual structure should be taken? Being the biggest does not entail being unique. Nevertheless, the exclusion postulate says nothing about this problematic *underdetermination of quale* (Krohn & Ostwald, 2017). Moreover, as we have noted at the end of section 2.1, what if the same Φ^{max}s are produced at the different spatio-temporal grains? When two equivalent Φ^{max}s are detected both from the level of neuronal units and from the level of cerebral lobes, the exclusion

20 The source of the combinatorial explosion is the combination principle. For example, as the calculation of φ needs be performed according to the combination principle, it must be carried out over all possible subsets of the candidate set and for each of those subsets over all possible purviews. Nevertheless, the main constraint is, that the number of unique bipartitions rises exponentially with the cardinality of the set (see Krohn & Ostwald, 2017, Appendix).

postulate cannot tell which level should be taken as the "locus" of conscious experience. At least at the current stage of the theory, IIT does not have any theoretical resource to deal with such issues.

3.4. From vector geometry to point geometry: the geometry of integrated information

IIT has always assessed integrated information in a geometrical manner. Nonetheless, several updates from IIT 1.0 to 3.0 brought some changes in the geometry of integrated information. Since IIT's central notions, such as concepts and conceptual structure, are closely related to the geometry of integrated information, a deeper analysis of why and how the geometry has been revised would be needed.

In earlier versions of IIT, the space for representing informational structures was called informal *qualia space*, the multidimensional space which has its axes for each possible state of the system[21] (Tononi, 2008; Balduzzi & Tononi, 2008, 2009). The geometrical shape expressing the informational structure was called *quale*. Q-arrows and points constitute the quale in qualia space: points represent actual repertoires specified by the system in a state when a certain connection–a set of causal connections–is added. Furthermore, q-arrows represent *informational relationships* between each actual repertoire specified by the added connection. Thus, the point "at the bottom" of the quale is a potential repertoire specified by the system in a state, when no connection is added (the "null set"). On the other hand, the point "at the top" of the quale is the actual repertoire, when all connections are added (the "full set"). By adding each connection from the potential repertoire, it is possible to analyze how much ei the system gains by each connection. This can provide detail about which connection

21 For example, when n is the total number of the system's elements and each element can have only two possible outputs, dimension of 2^n is required to constitute the qualia space which represents the system.

informationally contributes to the quale. All points connected by all q-arrows illustrate a geometrical figure like a "polytope" (see Balduzzi & Tononi, 2009, Figure 3).

The major point of the geometry of the earlier versions of IIT is that q-arrows are represented as vectors. Interpreting q-arrows as vectors, one can find many properties of the informational relationships constituting the quale: the length of the q-arrow represents how much ei is generated by adding a connection. For instance, the length of the q-arrow connecting "the bottom" and "the top" of the quale represents ei of the system in a state. Also, the direction of the q-arrow expresses *the particular way* how adding a connection sharpens repertoires. One of the most interesting properties of q-arrows, however, is *entanglement (γ)*: when a q-arrow cannot be decomposed into an exact vector sum of its sub-q-arrow, then it is considered as tangled. When the q-arrow is tangled, it means that there is integrated information gained by adding up the corresponding connection. The way how the q-arrow is tangled can be measured by vector calculation. The difference between the length of the q-arrow and that of the vector sum of its sub-q-arrows is quantified by γ. In this sense, entanglement represents how much information a q-arrow generates above and beyond its components. In an earlier version of IIT, the q-arrow with $\gamma>0$ was defined as a *concept*. Moreover, *complexes* could be defined by comparing γ of each concept; a concept with relatively high γ was called as a *mode*. Before IIT 3.0, these articulated analyses were available from the vector analysis of q-arrows (Tononi, 2008; Balduzzi & Tononi, 2008, 2009).

However, in IIT 3.0, such vector analysis is no longer available (Tononi, 2012; Oizumi, Albantakis, & Tononi, 2014). As explained in Section 3.3, the concept cannot be defined as an entangled q-arrow. Rather, it is defined as MICE plotted as a point in the concept space. Instead of the null set, the current version posits the "null concept," which is the unconstrained repertoire specified by the system of mechanisms when no mechanism of

the system is given. As explained in Section 3.3, one can calculate how much CI is generated by the mechanism by applying extended EMD. Thus, in concept space, the distance between two concepts does not capture how much ei is generated by adding a connection to the system in a state. Rather, it captures how much CI is generated by adding a mechanism to the system of mechanisms. In the current version of concept space, adding connections, specifying informational relationships between repertoires, and analyzing q-arrows cannot be found due to these differences. In sum, all analyses and notions grounded by vector calculus of q-arrows are not available in IIT 3.0. *Prima facie*, the geometry of integrated information appears to be simplified.

The transition from effective information to cause-effect information also affects the geometry of IIT. The transition in the notion of information doubles the concepts and space. Since the theory was based on ei, there were only the past repertoires. Therefore, just one space was required to represent concepts. In IIT 3.0, however, space must represent both the past and future states, because the theory is built on the notion of cei. As a result, there must be two repertoires that give φ^{max}: core causes and effects. Since the concepts cover not only the past but also the future repertories, space where the concepts are represented, and the geometrical structures made from concepts are doubled. In other words, the multidimensional space and MICS must cover both of the past and future. As the definition of information changes, almost everything in the geometry of IIT appears doubled: points, space, and geometrical structures. In this sense, the geometry of IIT seems to be rather complicated.

In a nutshell, the evolution of the geometry of integrated information has two sides. On the one hand, it has been simplified in that all the articulated vector analyses for q-arrows are not used anymore. On the other hand, however, it has been complicated in that every aspect of the geometrical approach should be counted twice. We believe that these double aspects of

the geometry of IIT are consequences of transitions to other central notions of the theory.

4. Theoretical Issues in IIT

Despite its scientifically interesting prospects, IIT also faces several theoretical extra-model problems. These problems concern IIT's principles, core concepts, and their possible consequences. Despite a few exceptions, many recent literatures are almost focused on particular technical issues (Kitazono, Kanai, & Oizumi, 2018; Hidaka & Oizumi, 2018).[22] This focus is fully understandable, since, without overcoming various technical barriers, there hardly would be empirical advances for IIT. However, theoretical problems deserve more attention, as it is theoretical considerations that enable us to judge whether or not the theory is worth pursuing in the first place. Despite such importance, theoretical issues have been largely overlooked in IIT debates, and relatively fewer studies have addressed this topic. Therefore, such theoretical problems IIT require closer analysis. Of note, while there might be many issues concerning IIT's theoretical aspects, in the present paper, we focus on three major problems that appear to raise serious questions about the plausibility of the theory.

4.1. Sophisticated panpsychism: Unjustified scientific authority

The first issue is that IIT embraces a form of *panpsychism.* Panpsychism has traditionally been ignored as full-fledged mysticism. The view that extremely simple organisms and even seemingly non-living things have "a small piece of mind" sounds counterintuitive enough. IIT, however,

22 Rare exceptions are (Bayne, 2018) and (Krohn & Ostwald, 2017). The former provides critical assessments of the axiomatic approach of IIT 3.0. The latter illustrates the important and disturbing conceptual issue of "magic cuts" which can violate IIT's fundament intuition: "the whole is more than the sum of its parts".

admits a variety of examples that could support a sophisticated sort of panpsychism. A representative case is that of photodiodes (Tononi, 2008; Oizumi, Albantakis, & Tononi, 2014). According to IIT, a photodiode, which is designed to react to various external stimulations only by lighting on and off, is "minimally conscious." This means that the photodiode has a minimal level of consciousness and a certain quality of experience as well. Nonetheless, the photodiode might be the last one we ascribe experience. It is difficult to believe that such a simple micro-mechanism could have a certain kind of consciousness. There is another example which appears to be the opposite of the photodiode case. Aaronson (2014a) has clearly shown that, if IIT is acceptable, then a lattice constituted by just connecting one kind of simple logic gates over and over could have a high value of Φ. According to Aaronson's (2014a) description, XOR gates arranged in a 2-D square grid would be conscious. Much astonishing result is that such increasing of Φ is proportional to the length and breadth of the grid.[23] Therefore, by a simple recursive procedure of connecting more XOR gates, there could always be a huge physical lattice which is more conscious than a normal human being! Though this is truly unbelievable, IIT allows these examples.

If the photodiode refers to the micro-case of panpsychism, the lattice could be its macro-case. The problem is that those simple and non-organic things' being conscious is so counterintuitive that it would rather be easier to take it as a *counterexample* than as evidence. If IIT predicts that those simple systems that are unconscious could be conscious, at least for many, such prediction itself would be enough to present a *reductio ad absurdum* against IIT. Hence, the charge of panpsychism should be taken seriously in deciding whether or not IIT is theoretically plausible. If it is certain that

23 To be fair, Aaronson's calculation was based on IIT 2.0 so that it cannot be directly applied to the current version. Unlike IIT 2.0, IIT 3.0 does not require procedure of normalization.

no simple object such as a photodiode or an XOR grid is conscious and therefore panpsychism is wrong, IIT must be wrong too.

What makes this issue more problematic is that the founders of IIT are seemingly undaunted by those critiques above. According to the founders, IIT does not just allow diverse panpsychistic cases; instead, IIT actively argues for those counterintuitive cases by demonstrating them in a detailed manner. Tononi's (2014) reply shows his confidence that all those counterintuitive cases are evidence for IIT. Tononi (2014) emphasizes that when science and common sense conflicts with each other, it is always science that takes priority. As stated in Tononi (2014), this is the primary reason why we should count those hard-to-swallow examples as evidence. The history of science is full of reversions of the common sense by innovative scientific discoveries. Since IIT is a scientific theory, the fact that IIT produces several counterintuitive predictions cannot be a strong reason to reject it. Rather, in Tononi's (2014) view, it is our widely entrenched intuition that must be corrected. Said differently, IIT might be on the edge of "scientific revolution," and Tononi might be, following Aaronson's witty phrase (Aaronson, 2014b), "the Copernicus-of-consciousness."

Nonetheless, Tononi's (2014) reply could be objected in several ways. First, it is not obvious at all that IIT is able to claim its priority over common sense or intuition. Even if it is true that science tends to override culturally and historically widespread intuitions, the question remains whether IIT has any right to do so. Technically, not all hypotheses of science can have the right to correct popular intuitions. In Kuhnian terms, only the so-called "normal science," which has successfully secured, well-established methodologies, exemplars, problem-solving procedures, basic beliefs and values shared by members of the scientist society, can argue for its right over common sense and intuition (Kuhn, 1962). However, it seems undeniably clear that, in the current stage, IIT cannot be such a normal science of consciousness. For now, it is nothing more than an interesting

working hypothesis that should wait for rigorous examination from the current scientist society. Moreover, as repeatedly pointed out in Section 3, IIT suffers from some technical issues preventing empirical experiments and practical applications. No direct evidence has been obtained by empirical studies conducted on real physical systems. Despite the growing body of empirical studies resting on the IIT framework, no IIT theorist has been able to apply the pure IIT 3.0 to neural data such as brain signals.

Given its present status in the field, IIT appears to be unfit to serve as a hypothesis of normal science. Rather, IIT is more likely to be something in between "pre-science" and normal science, which might be one possible candidate of a "paradigm shift" in the field of consciousness studies. Then, IIT's panpsychistic predictions cannot be before our general intuitions about consciousness. A heavy burden of proof is still on the side of IIT, and our anti-panpsychistic intuition should be taken as default. Aaronson's (2014b) comment reveals this situation: "The anti-common-sense view gets all its force by *pretending* that we're in a relatively late stage of research—namely, the stage of taking an agreed-upon scientific definition of consciousness, and applying it to test our intuitions—rather than in an extremely early stage, of agreeing on what the word "consciousness" is even supposed to mean(italics added)".

The problematic implication of sophisticated panpsychism does not lie only in the conflict with the strong intuitions, which is external to IIT. It also lies in the logical development of the structure of the theory, which is internal to IIT. It is the most original and unique feature of IIT that the theory starts from some phenomenological axioms. The problem is that the axioms are taken for granted in IIT. They are assumed to be self-evident. However, taking something for granted or assuming it to be self-evident is just another way of accepting it as intuitive. In this sense, it is IIT itself that strongly depends on a set of intuitions. IIT is fundamentally *grounded* on

several phenomenological intuitions.[24] Hence, if IIT allows panpsychistic cases and denies opposing intuitions, *a charge of double standards* could be raised. On the one hand, IIT strongly holds some intuitions by calling them "phenomenological axioms." On the other hand, it easily dismisses other intuitions by treating them unscientific common sense. Nonetheless, how can IIT justify this selective adoption of intuitions? Why does it adopt one group of intuitions but reject another? If axioms of IIT are considered as a significant type of phenomenological intuition concerning what consciousness *is*, anti-panpsychistic intuitions should also be taken to be equally important phenomenological insights about what consciousness is not. At least in the current version of IIT, we cannot find any principled reason to take axioms for granted and to reject other intuitions about consciousness. Once IIT wants to deny anti-panpsychistic intuitions as prejudices of scientifically unenlightened non-specialists, it should do the

24 To this matter of grounding IIT, one reviewer has raised an interesting point. The reviewer predicted that "defendants of IIT would argue that the set of axioms is qualitatively different from anti-panpsychist intuitions in that they are not only self-evident but also directly accessible from a first-person perspective". While this might be true, this reply seems to raise another issue about the direct accessibility of consciousness and its fundament properties, on which phenomenological axioms are about. This 'direct accessibility from the first-person point of view' has usually been discussed under the title of *introspection.* In order to claim that introspection lends further support to the phenomenological axioms, one must first prove that such introspection is significantly reliable enough to have some evidential force. However, it is controversial if introspection is significantly reliable; rather, a growing number of empirical studies suggest that introspection is not a reliable source of evidence. Once this point is taken, the alleged qualitative difference between anti-panpsychistic intuition supporting common sense and phenomenological axioms grounding IIT becomes doubtable. Though the reliability of introspection deserves deeper analysis, in the current context, raising doubt against introspection is enough to elaborate our argument by blurring the difference between anti-panpsychistic and phenomenological intuition. For a thorough critical assessment of the reliability of introspection, see (Schwitzgebel, 2008, 2013). Smithies and Stoljar (2012) also present ample philosophical arguments for or against the special nature of introspection.

same thing with its underlying intuitions. However, what such denial of its axioms amounts to is just a self-refutation. Therefore, without providing further reason to take its axioms and ignore anti-panpsychistic intuitions, IIT cannot be free from its charge of double standards of contrasting intuitions.

To sum up, sophisticated panpsychism implied by IIT threatens IIT itself in two ways. First, considering IIT's premature status, the panpsychistic charge gives a very good reason to reject IIT. As long as no strong evidence is provided, panpsychism alone could suffice not to believe IIT. Also, it raises the charge of double standards to seemingly respectable intuitions. Being fundamentally founded by "phenomenological axioms," it is difficult for IIT to dismiss opposing intuitions.

4.2. Fading and Dancing qualia: Radical dissociation between experience and cognition

The second issue with IIT is that as Cerullo (2015) has pointed out, IIT faces the Fading and Dancing qualia arguments.[25] Fading and Dancing qualia refer to thought experiments designed by David Chalmers (1996). As suggested by their names, Fading qualia describe an imaginable situation where qualia become more and more eroded. Dancing qualia show another scenario that the whole qualia are replaced by totally different qualia. Their purpose is to show that, in our natural world, any attempt to detach experience from the functional organization of a system would face extremely counterintuitive consequences. Despite the richness of detail, in the context of IIT, the relevant point is simple: IIT appears to entail anti-functionalism or anti-computationalism so that it commits to a possibility which Fading and Dancing qualia rule out.

25 Although Cerullo (2015) highlights the point, he does not provide a specific description or analysis in his work. By contrast, Shanahan (2015) provides a more clear and comprehensive analysis. Both of them concerns upon the problem of Fading and Dancing qualia anyhow.

Fading qualia start with the assumption of the physical system and its functional organization in our world. Since the functional organization is a matter of abstraction, it must be fixed how far the organization should be grained. In fading qualia, functional organizations are supposed to be sufficiently fine-grained to fix physical systems' *behavioral capacities.* Following this assumption, if two physical systems share the same functional organization, all their behaviors must be identical. Another assumption is *multiple realizations* without experience. It is assumed that there are multiple kinds of materials in implementing one organization, but only some of them support the phenomenal qualities of experience accompanied by the organization, while others do not.

Now let us imagine that neurons realize the functional organization of Mary's brain. Then, Mary sees a ripe tomato and feels a visually red feeling. In her brain, maybe somewhere in her visual cortex, there is a neural correlate of that red quale. Then, something strange happens. The neurons composing her neural correlate of phenomenal redness are now substituted by silicon chips one by one. Given multiple realizations, this replacement must be possible. The crucial point is that, although those chips are perfect functional equivalents of Mary's neurons, they do not support any quale at all. Therefore, if Mary's neurons get substituted by silicon chips, her vividly red experience should become murkier, and eventually will disappear. The problem is that Mary's functional organization never undergoes any change, despite of the gradual qualitative change of her experience. She would still manifest the same bodily and verbal behaviors as before. Moreover, considering that her brain function is perfectly the same, it is reasonable to think that her *cognitive states* are also the same as before. If cognitive states of Mary, such as her judgments or beliefs about the experience, do not remain intact and change following the eroding visual experience, such cognitive states would radically come apart from the functional organization of Mary' brain. Nothing in the functional organization

would correspond to the change of cognitive states. Chalmers argues that this kind of dissociation is highly unlikely, by saying, "If such a major change in cognitive contents were not mirrored in a change in a functional organization, cognition would float free of internal functioning like a *disembodied Cartesian mind*." (Chalmers, 1996, p. 258, italics added) This is why he claims that "[t]here is simply no room in the system for any new beliefs to be formed, … [u]nless one is a dualist of a very strong variety." (Chalmers, 1996, p. 258) As free-floating, disembodied cognitive states are deeply problematic and counterintuitive: it is safe to assume that cognitive states do not undergo any changes.[26] As a result, Mary neither notices nor is aware of anything. This is fading qualia in a nutshell. It seems highly unlikely that such a situation could occur in our world. What is worse is that Mary is perfectly rational and functional in every other aspect, except for her beliefs about her visual experience. She is not pathological or deeply confused. Nevertheless, she suffers somewhat systematic errors concerning her experience. Whenever a substitution occurs, she forms the wrong belief that she still sees the red tomato. This systematic error of rational subject is hardly acceptable in our natural world.

Dancing qualia is another version of fading qualia. In fading qualia, the phenomenal aspect of the experience gets gradually eroded and ends up to none. In dancing qualia, however, the phenomenal aspect does not vanish. Instead, it keeps changing. Mary does not suffer a gradual neuron-silicon replacement. Nonetheless, she has a certain neuroprosthetic device, which functions identically to her natural neural correlates of the reddish quale. This time, despite its function, the device does not support the reddish quale. Suppose that it grounds a blue quale instead. And there is a switch that alters Mary's neural correlate to the device. Then, what would happen

26 One of the reviewers advised that there should be more rationales to claim that cognitive states are fixed under the gradual replacement. For more on the debate, see (Chalmers, 1996, p. 247-274).

if someone turns the switch on? *Ex hypothesi*, Mary's visual experience will suddenly become blue-like. If the switch turns off, the opposite will happen. Hence, as someone turns the switch on and off, Mary's visual quale will dance back and forth! The trouble is that Mary would not be able to notice any change in her visual field. Since the device is the perfect functional duplicate of Mary's neural correlates, the functional organization of her brain remains the same. As in fading qualia, Mary's cognitive states would be intact, regardless of the change of the phenomenal aspect of her visual experience. If so, Mary would neither notice nor be aware of any change, even if visual qualia are dancing "in front of her eye"! For the same reason as in fading qualia, it appears that this consequence must be rejected.[27]

The relevant point in the context of IIT is that IIT essentially allows these implausible cases. Fading and Dancing qualia are possible only on the assumption that there could be functionally identical, but phenomenally different systems. In the IIT framework, neurons and silicon chips, neural correlates and the neuroprosthetic device could be such systems. The only way to prevail the unwelcomed consequences appears to be denying the possibility of the functionally identical, but phenomenally different systems. IIT, however, does not and even cannot deny that possibility. According to IIT 3.0, even if two physical systems perfectly share their functions, they can be different in Φ^{max} they produce. Considering that maximally integrated conceptual information is the experience in IIT, the claim that function and Φ^{max} can come apart implies *anti-functionalism or anti-computationalism* about consciousness. There could be *zombie systems* which perform the same as conscious systems but do not have any experience at all. This is not just speculation; indeed, Oizumi, Albantakis, and Tononi (2014) design such a zombie system and demonstrate how it

27 The Fading and Dancing qualia arguments are also related with computational approach in cognitive science. Chalmers uses the arguments to defend his computationalism and other philosophers and scientists raises several objections against the view. See (Chalmers, 2011; Harnad, 2012)

works (Ibid., Fig 21). If a zombie system is possible, there is no reason not to believe silicon chips in fading qualia or neuroprosthetics in dancing qualia. Then, IIT should accept those unacceptable consequences anyway.

It is IIT's anti-functionalism that opens the door to Fading and Dancing qualia. In front of the implausible results of Fading and Dancing qualia, there are only two logical ways for IIT to reply: to dodge the bullet or to bite it. Nonetheless, none of the two appears to be available without significant revisions of the theory. On the one hand, if IIT wants to dodge the bullet, it must show how Mary could notice the change in her visual experience, even if her brain functions remain the same. It is highly likely that, if there is no difference in the brain functions, the same will apply to information processing. In IIT as a paradigm of cognitive science and artificial intelligence, it is widely accepted that there should be corresponding activities of information processing to notice or be aware of something. However, by the assumption of functional identity, Mary cannot have any new information processing corresponding to the change of quale. Then, how can Mary notice or be aware of the experiential change? On the other hand, if IIT tries to bite the bullet, all the debates concerning the charge of double standards resurface again. IIT cannot merely say, "Though being counterintuitive, nonethelessly, it's true." IIT is scientifically so premature that it is not in a position to override strong intuitions in the name of science. Furthermore, since IIT itself takes some intuitions as primitive, it cannot easily dismiss other intuition as ungrounded. In one way or another, it seems difficult for IIT to defy the intuition that *the radical dissociation between cognition and experience* is impossible. In one way or another, IIT can neither dodge nor bite the bullet of Fading and Dancing qualia.

All in all, IIT cannot deal with Fading and Dancing qualia. Holding anti-functionalism about consciousness, IIT does not have theoretical resources to explain how the system which functionally remains identical could

notice its phenomenal changes. On the other hand, accepting the possibility of unnoticeable phenomenal change is extremely counterintuitive to that, if IIT allows such notion, many will reject IIT. As in the panpsychism debate, due to its dependence on intuitions, IIT cannot merely dismiss the intuition that a rational and functioning system must be able to be aware of its experiential changes. Either way, IIT faces serious troubles.

4.3. The paradox of certainty: Loss of certainty undermines existence

In Section 4.2, we argued that, although the empirical possibility of radical experience-cognition dissociation causes a serious counterintuitive consequence, IIT cannot dodge this consequence. In this section, we attempt to show that such radical experience-cognition dissociation causes another problem: *the loss of certainty about consciousness.* We believe that this loss of certainty can undermine the very foundation of IIT: *the existence of consciousness.*

We, or at least many of us, appears to be *certain* about our consciousness. Our consciousness might be the only thing we can be certain. However, the argument from Fading and Dancing qualia shows that our phenomenal beliefs or judgments can be detached from our consciousness even when we are fully alert and attended. If this is the case, we might be suffering Fading and Dancing qualia as well. That is, we might be like Mary who cannot be aware of the absence of her visual consciousness. If so, even if we strongly believe or take for granted that we are conscious here and now, it is possible that we are not. As Descartes doubted, an omnipotent demon might manipulate our perceptual experience to make us believe the existence of the external world, even if there is no such world. Similarly, something might control our cognitive system to make us believe the existence of our experience, even if there is no such thing as experience at all. Then, how can we be so sure that we are conscious here and now? In other words, is there any guarantee that we are not *deluded zombies*

who think that they are conscious if experience and cognition about the experience can come radically apart? It is clear that the radical experience-cognition dissociation deprives us of the certainty of consciousness. And if IIT allows the dissociation, it cannot secure the certainty of consciousness.

Some might deny the certainty of consciousness. While the certainty of our own experience is seemingly undeniable, whether or not we are certain about our experience is surely debatable. Nevertheless, it appears that IIT cannot easily deny the certainty of consciousness, because the theory appears to be *grounded* in it: the first phenomenological axiom states that consciousness exists. Furthermore, this existence of conscious experience is supposed to be certain. Indeed, it is argued that consciousness is certain when Tononi (2012) paraphrases Descartes' *cogito ergo sum*: "I experience therefore I am" (p. 296). The very starting point of IIT, the existence axiom, necessarily requires the certainty of consciousness. If we are not certain about our consciousness, why should we struggle for a scientific theory of consciousness?

Therefore, the possibility of the radical experience-cognition dissociation provides a somewhat delightful and disturbing paradox against IIT: If IIT is true, radical experience-cognition dissociation is possible. If so, we cannot be certain about our consciousness. If we cannot be certain about our consciousness, IIT cannot get off the ground. Therefore, if IIT is true, there is no reason to suppose that it is true. We call this argument *the paradox of certainty*. IIT appears to require and reject the certainty of consciousness simultaneously.

It seems that the only possible reply from IIT would deny the empirical possibility of the radical experience-cognition dissociation. However, as we have seen in Section 4.2, the problem is that, at least in the current version of the theory, it is difficult to find any reason for such denial. Because IIT argues for the functional zombie system, it is doubtable that IIT can deny such a possibility. We cannot find any consideration about how experience

affects beliefs or judgments, and vice versa in IIT. While IIT appears to have a great deal with how experience is generated from its physical substrate, it does not provide much insight into how the subject can be aware of that generated experience. Said differently, IIT is blind to the question of how we can secure *self-knowledge or metacognition* about our own experience. This is the topic of the last section of this paper.

4.4. Metacognitive accessibility: Missing link in IIT

What is the main source of the theoretical problems mentioned thus far? We think the culprit here is disregarding cognitive aspect of consciousness.[28] In IIT, the explanation of how a subject could cognitively access the experience is absent. IIT never takes account of *metacognition* in explaining consciousness, and we believe that it is this neglect of metacognition that generates all theoretical problems IIT faces.

Due to its ignorance of metacognition of consciousness, IIT can ascribe consciousness to simple systems lacking metacognitive mechanisms, such as photodiodes or logic grids. Though photodiodes and logic grids produce integrated information, it is highly unlikely that these simple physical systems are equipped with metacognitive mechanisms. Given that they lack metacognition, those systems do not, and even cannot, have cognitive access to integrated information of their own. There is no photodiodes and logic grids' metacognition of their integrated information. Under the IIT framework, this metacognitive inaccessibility implies that photodiodes and logic grids cannot know or be aware of their consciousness. While they are conscious, they cannot know that they are conscious! However, this lack of metacognition and its strange consequence do not prevent IIT to ascribe consciousness to simple systems, as it does not concern metacognitive

28 Cerullo (2015) makes a similar point. After distinguishing *incognitive* and *cognitive* consciousness, he argues that IIT only deals with incognitive one, which is tantamount to consciousness without subject.

access to consciousness at all.

Furthermore, since IIT appears to neglect how metacognition and experience could be associated, it allows the radical dissociation between metacognition and experience. Such dissociation is shown by Fading and Dancing qualia and ultimately results in the paradox of certainty. In Fading and Dancing qualia, unlike in the panpsychistic cases, the system has metacognitive access to integrated information it produces. That is, Mary has a metacognitive belief about her visual experience. The problem is that her metacognitive access systematically produces wrong beliefs about her own experience. In fading qualia, Mary is usually right about what she sees. However, as soon as the process of neuron-to-silicon replacement begins, Mary starts to have wrong beliefs about what she sees. In dancing qualia, whenever the switch turns on, Mary becomes wrong about her visual experience. In both cases, Mary's being wrong is very systematic in that it strongly correlates with the replacement. Mary's systematically being wrong indicates that her metacognitive access to her visual experience systemically results in wrong beliefs. However, since there is no consideration about how the system metacognitively accesses its own experience in IIT, it cannot help but allow the absurdities of Fading and Dancing qualia. Also, once the radical experience-cognition dissociation is admitted as possible, there appears to be no way to eschew the paradox of certainty.

Given the tight relationship between experience and cognitive access, IIT's neglect of metacognition is somewhat surprising. Phenomenologically, there appears to be a close or even constitutive relation between metacognition and experience. Despite philosophical debates surrounding the distinction between phenomenal vs. access consciousness (Block, 1995, 2007), we believe that there could be experience without actual metacognitive access. Nevertheless, this does not mean that there could be an experience that cannot be metacognitively accessible. It sounds absurd and even unintelligible that conscious experience is absolutely out of our

range of metacognition. Such experience must be a *conscious experience we cannot be conscious of*, which should be defined as unconscious by its nature. Hence, it appears that metacognitive *accessibility*, not actual metacognitive *access*, is necessarily involved in having consciousness. That is, metacognitive accessibility is a necessary condition for something to be a conscious experience.[29]

Therefore, we argue that any scientific theory of consciousness must take account of the metacognitive accessibility of consciousness. However, no matter which version it may take, IIT does not seem to consider why and how metacognitive accessibility must be taken into account when it comes to explaining conscious experience. Accordingly, we strongly suggest that the first step to deal with the theoretical problems mentioned so far is introducing metacognitive accessibility in the IIT framework. Phenomenological axioms, ontological postulates, and mathematical models of IIT should be revised to reflect the necessary connection between metacognitive accessibility and consciousness. Once we can successfully assimilate metacognition into IIT, we could have a better version of the theory, which would deserve to be called 'IIT 4.0.'

---

29 For a similar point, see Chalmers (1997) who argues against Block (1995) that, even when there is *phenomenal consciousness* (P-con) without *access consciousness* (A-con), it does not mean that there is not accessible consciousness. According to Chalmers (1997), once A-con is defined in terms of availability for global control, P-con always goes along with A-con. Since global availability requires only accessibility, the original notion of A-con should be modified from access consciousness to accessible consciousness. Our suggestion here could be taken as claiming that, if an experience is phenomenally conscious, it must be accessibly conscious. It is worth noting that this transition from access to accessibility is what distinguishes Chalmers (1997) and us from those who follow *Higher Order Theory of consciousness* (HOT) (Rosenthal, 1986, 2005). In HOT, for a mental state to be conscious, it must be actually accessed by a higher order state. Our suggestion, however, does not demand actual higher order, metacognitive access. All that required is that the state must be metacognitively accessible. No actual higher order state needs to be there.

5. Conclusion

IIT has been the center of the debates surrounding the science of consciousness. Many of those who are engaged in the field displayed interest in the theory, and some raised serious doubts and criticisms. It is worthwhile to assess what IIT is about and why it is controversial. In this paper, we have critically examined the theoretical evolution and related issues of IIT. We have introduced basic concepts, which might be considered as the core of IIT. Both IIT's explanatory power and limits appear to be already embedded in its core concepts. We have also described how the theory has been updated throughout the last decade. In some aspects, those major transitions can be thought of as progress. However, in other aspects, some of the issues were worsened, and even new problems emerged. Specifically, the principled part of the framework of IIT, its phenomenological axioms, and ontological postulates raise serious questions about the scientific status of the theory, the possibility of radical dissociation between experience and cognition, and the logical structure of the theory. We have suggested that focusing on our ability to access our own experience through metacognition might be one way to deal with these theoretical issues. The cognitive relationship between metacognition and consciousness might push IIT one step forward in becoming a normal science of consciousness.[30]

Author contribution

HP wrote sections 2 and 3; KM wrote sections 1, 4 and 5; All authors reviewed the manuscript

30 We gratefully thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

Reference

Aaronson, S. (2014a). *Why I Am Not an Integrated Information Theorist (Or, The Unconscious Expander).* Retrieved On June 2018, From The Website: Https://Www.Scottaaronson.Com/Blog/?P=1799.

Aaronson, S. (2014b). *Giulio Tononi and Me: A Phi-Nal Exchange.* Retrieved On June 2018, From The Website: Https://Www.Scottaaronson.Com/Blog/?P=1823.

Balduzzi, D., & Tononi, G. (2008). *Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework.* Plos Comput Biol 4(6): E1000091. Https://Doi.Org/10.1371/Journal.Pcbi.1000091.

Balduzzi, D., & Tononi, G. (2009). *Qualia: The Geometry of Integrated Information.* Plos Comput Biol 5(8): E1000462. Https://Doi.Org/10.1371/Journal.Pcbi.1000462.

Barrett, A. B., & Seth, A. (2011). *Practical Measures of Integrated Information for Time-Series Data.* Plos Comput Biol 7(1): E1001052. Https://Doi.Org/10.1371/Journal.Pcbi.1001052.

Bateson, G. (1972). *Step to Ecology of Mind.* University of Chicago Press

Bayne, T. (2018). *On The Axiomatic Foundations of the Integrated Information Theory of Consciousness.* Neuroscience of Consciousness, Volume 2018, Issue 1, Niy007, Https://Doi.Org/10.1093/Nc/Niy007.

Block, N. J. (1995). *On A Confusion About the Function of Consciousness*. Behavioral and Brain Sciences 18: 227-247.

Block, N. J. (2007). *Consciousness, Accessibility, And The Mesh Between Psychology and Neuroscience.* Behavioral and Brain Sciences. 30: (5-6):481-99; Discussion 499-548.

Cerullo, M. A. (2015). *The Problem with Phi: A Critique of Integrated Information Theory.* Plos Comput Biol 11(9): E1004286. Https://Doi.Org/10.1371/Journal.Pcbi.1004286.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

Chalmers, D. J. (1997). *Availability: The Cognitive Basis of Consciousness?*. Behavioral and Brain Sciences 20: 148-149.

Chalmers, D. J. (2011). *A Computational Foundation for The Study of Cognition.* Journal of Cognitive Science 12: 323-357.

Harnad, S. (2012). *The Causal Topography of Cognition.* Journal of Cognitive Science 13: 181-196.

Hidaka, S., & Oizumi, M. (2018). *Fast and Exact Search for The Partition with Minimal Information Loss.* PLOS ONE 13(9): E0201126. Https://Doi.Org/10.1371/Journal.Pone.0201126.

Horgan, J. (2015). *Can Integrated Information Theory Explains Consciousness?.* Retrieved On June 2018, From The Website: Https://Blogs.Scientificamerican.Com/Cross-Check/Can-Integrated-Information-Theory-Explain-Consciousness.

Kitazono, J., Kanai, R., & Oizumi, M. (2018). *Efficient Algorithms for Searching the Minimum Information Partition in Integrated Information Theory.* Entropy, 20(3), 173; Doi:10.3390/E20030173.

Krohn, S. & Ostwald, D. (2017). *Computing Integrated Information.* Neuroscience of Consciousness. Volume 2017, Issue 1, Nix017, Https://Doi.Org/10.1093/Nc/Nix017.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press

Marshall, W., Gomez-Ramirez, J., & Tononi, G. (2016). *Integrated Information and State Differentiation.* Frontiers in Psychology, 7, 926. Http://Doi.Org/10.3389/Fpsyg.2016.00926.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). *From The Phenomenology to The Mechanisms of Consciousness: Integrated Information Theory 3.0.* Plos Comput Biol 10(5): E1003588. Https://Doi.Org/10.1371/Journal.Pcbi.1003588.

Oizumi, M., Amari, S., Yanagawa, T., Fujii, N., Tsuchiya, N. (2016). *Measuring Integrated Information from The Decoding Perspective.* Plos Comput Biol 12(1): E1004654. Https://Doi.Org/10.1371/Journal.Pcbi.1004654.

Rosenthal, D. (1986). *Two Concepts of Consciousness.* Philosophical Studies. 49: 329–359.

Rosenthal, D. (2005). *Consciousness and Mind.* Oxford: Oxford University Press.

Schwitzgebel, E. (2008). *The Unreliability of Naive Introspection.* Philosophical Review 117: 245-273.

Schwitzgebel, E. (2013). *Perplexities of Consciousness.* A Bradford Book; Reprint Edition.

Schwitzgebel, E. (2012). *Why Tononi Should Think That the United States Is Conscious*. Retrieved On June 2018, From The Website: Http://Schwitzsplinters.Blogspot.Kr/2012/03/Why-Tononi-Should-Think-That-

United.Html.

Schwitzgebel, E. (2014). *Tononi's Exclusion Postulate Would Make Consciousness (Nearly) Irrelevant*. Retrieved On June 2018, From The Website: Http://Schwitzsplinters.Blogspot.Kr/2014/07/Tononis-Exclusion-Postulate-Would-Make.Html.

Smithies, D. & Stoljar, D. (2012). (Eds.) *Introspection and Consciousness*. Oxford University Press.

Shanahan, M. (2015). *Ascribing Consciousness to Artificial Intelligence*. Arxiv:1504.05696v2 [Cs.AI].

Tegmark, M. (2016). *Improved Measures of Integrated Information*. Plos Comput Biol 12(11): E1005123. Https://Doi.Org/10.1371/Journal.Pcbi.1005123.

Tegmark, M. (2017). *Life 3.0: Being Human in The Age of Artificial Intelligence*. Penguin UK.

Tononi, G. (2001). *Information Measures for Conscious Experience*. Archives Italiennes De Biologie, 139:367–71.

Tononi, G. (2004). *An Information Integration Theory of Consciousness*. BMC Neuroscience. Https://Doi.Org/10.1186/1471-2202-5-42.

Tononi, G. (2008). *Consciousness as Integrated Information: A Provisional Manifesto*. The Biological Bulletin 215, No. 3: 216-242. Https://Doi.Org/10.2307/25470707 PMID: 19098144.

Tononi, G. (2010). *Information Integration: Its Relevance to Brain Function and Consciousness*. Archives Italiennes De Biologie, 148: 299-322.

Tononi, G. (2012). *Integrated Information Theory of Consciousness: An Updated Account*. Archives Italiennes De Biologie, 150: 290-326.

Tononi, G. (2014). *Why Scott Should Stare at A Blank Wall and Reconsider (Or, The Conscious Grid)*. In Aaronson, S. (2014a).

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). *Integrated Information Theory: From Consciousness to Its Physical Substrate*. Nature Reviews Neuroscience Volume 17, 450-461, Doi:10.1038/Nrn.2016.44.

Tononi, G. & Koch, C. (2015). *Consciousness: Here, There, And Everywhere?*. Phil. Trans. R. Soc. B 2015 370 20140167; DOI: 10.1098/Rstb.2014.0167.

Virgil, G. & Koch, C. (2014). *Quantifying Synergistic Mutual Information*. In Prokopenko M. (Eds.), Guided Self-Organization: Inception. Emergence, Complexity and Computation, 159-189. Springer.