Cardinality Criterion for the Integrated Information Theory

Ahmet Oğuz Yemenici | *Bilkent University*

Abstract

The Integrated Information Theory of Consciousness (IIT) is one of the most prominent theories in neuroscience. In this paper I offer a mathematical criterion to test out the theory's soundness. The theory (IIT 3.0) posits a mathematical object that can be generated in neural systems, called: *maximally irreducible conceptual structure* (MICS), while the latest version (IIT 4.0) employes a different mathematical structure called Φ-structures. Both of these versions later claim that every instance of a phenomenal conscious experience is *identical* to a MICS/Φ- structure. That is to say, the distinguishing features of a conscious experience from others, are in fact, nothing over and above the distinguishing features of a given MICS/Φ-structure to other MICSs/Φ-structures. This entails that the number of all possible and actual conscious mental states of a person is equal to the number of all possible and actual MICSs/Φ-structures that can be generated by that person's neural system. I argue that the cardinality of the class of all possible conscious mental states for a person, is at least countably infinite. Then on that basis, I argue that for IIT to be true, it needs to predict that the cardinality of the class of all MICSs/Φ-structures of a person should be greater or equal to countably infinite.

Introduction

IIT's main program is to detect the essential properties of every conscious mental state and axiomatize them (Tononi, 2015). The theory later turns these axioms into mathematical postulates to define a kind of mathematical construct that has the characteristics of conscious mental states (Albantakis et al., 2023, p. 3-5). One of the key claims of the theory states that the physical substrates which realize this type of mathematical construct generates consciousness (Albantakis et al., 2023, p. 5). In that way, theory hopes to achieve a quantized model for conscious experiences.

The essential properties that are axiomatized by IIT, to be a conscious mental state are: *existence*, *intrinsicality*, *information*, *integration*, *exclusion* and *composition* (Albantakis et al., 2023, p. 5). The *existence* axiom simply states that the conscious system exists. In some sense, it should be able to both affect and be affected by objects. The *intrinsicality* axiom is about the domain of this cause-effect potency, that is to say, the conscious system should be able to form cause-effect relations *intrinsically*. The next axiom is the axiom of *integration.* The axiom holds that conscious experiences are unitary. In a sense, it states that there can be only one conscious state at a time. (i.e., the composition of the experience cannot be divided into smaller pieces that are also conscious states themselves.) The *information* axiom states that every conscious mental state is unique. There is always a distinguishing feature between two seemingly similar mental states. The *exclusion* axiom is about the *borders* of conscious experiences. In this lexicon, the axiom states that a conscious experience has a specific content and nothing more. And the last axiom, *composition*, concerns simply with the fact that there is a structure to conscious experiences. For example, in spatial experiences there is a left and a right side of the sight.[1]

The key aspect of the IIT for this paper is the identity claims of the theory. The former version of the theory, IIT 3.0, holds that every conscious experience or *quale*[2] is identical to a mathematical structure called *maximally irreducible conceptual structure* (MICS), also named as *quale sensu lato* (Tononi, 2015).

1 The six axioms and their correspondent postulates hold relatively small importance to the content of this paper. Therefore, this brief paragraph will be all that is dedicated to them. More detail can be found in the bibliography.

2 *Quale* denotes spesific and individual mental states (e.g., tasting Döner, having paizzn or seeing a red apple for the first time) (Tye, 2021).

MICSs are mathematical forms which qualifies physical substrates to be realizers of conscious experiences. The physical substrates in question might be neural systems or any other physical entity that can generate MICS. For example, the particular conscious mental state of person *A*, tasting döner while listening traffic noise at 7:34 pm. 14 July 2023, is identical to a specific MICS that is generated by person *A*'s neural system. If there was even a minute change that makes relevance to the experience, maybe less seasoning in the döner, then the conscious experience would be different as well as the MICS. It is important to note that the identity relation IIT 3.0 draws, is between conscious experiences (*qualia*) and MICSs. The identity is not between the conscious experience and the physical substrate of the conscious agent (Tononi, 2015). Thus, the theory commits that a particular döner tasting experience and the brain state which generates the relevant MICS, are not identical; the experience and the MICS's itself are.[3]

The latest version of the theory, IIT 4.0, holds *Φ-structures* to be identical with conscious experiences, not MICSs (Albantakis et al., 2023, p. 29). This version names the physical substrates that generate the Φ-structures as, *complexes*. In other terms, *complexes* which generate certain kinds of cause-effect structures are Φ-structures. IIT 4.0's notion of Φ- structures are similar to the IIT 3.0's notion of MICS, for the fact that they are both held identical to the conscious experiences. However, IIT 4.0 has a nuance in its identity claim: "IIT proposes an explanatory identity: every property of an experience is accounted for in full by the physical properties of the Φ-structure…" (Albantakis et al., 2023, p. 6). That is to say, a particular Φ-structure and its correspondent conscious mental state are identical in the sense that, every property of the mental state is captured and accounted by a correspondent property of the given Φ-structure.

Metaphysical Implications

IIT 3.0 endorses a metaphysically necessary connection between conscious mental states and their correspondent MICSs, for they are numerically identical. In other terms, a specific conscious mental state *X* cannot fail to exist as long as its correspondent MICS exists and vice versa. This picture also entails there to be a modal relation between these two:

[3] Due to this reason, IIT bypasses some objections that arise from holding mental states identical to brain states

(1) The number of all possible and actual conscious mental states of a person *A* that has the neural system *N*, is equal to the number of all possible and actual MICSs that can be and are generated by that neural system *N*.

The above modal claim is similar to this: The number of all possible and actual water molecules is equal to the number of all possible and actual H_2O molecules, for they are identical. The same is also the case for IIT's claim. This fact makes these two classes' cardinality equal.

IIT 4.0 on the other hand, offers a different type of identity claim: explanatory identity (Albantakis et al., 2023, p. 5-6). I take it that, the IIT 4.0 theorists want to subscribe a more moderate kind of identity that is not as strong as numerical identity. In the context of IIT 4.0, explanatory type identity seems to propose a one-to-one mapping between features of a conscious mental state and its unique Φ-structure. The said identity relation seems to resemble with a map of a territory. However, the map is so detailed that the map and the territory has the same complexity and details. It's as if, for each property of the territory, there is a property in the map. To illustrate it, let's say that conscious mental state *K* has the properties $P1 = \{F1, \dots Fn\}$ and correspondent Φ-structure has the properties $P2 = \{G1 \dots Gn\}$. IIT 4.0 does not want to endorse an identity between these two classes that would entail properties to be identical respectively such as $F1 = G1$, $F2 = G2 \dots Fn = Gn$. Rather, it employes a function to link class P1 to class P2, so that every property of conscious state *K*, will be accounted by correspondent properties in the class P2. Just like in IIT 3.0, we can make a modal claim of this sort for the IIT 4.0 as well:

(2) The number of all possible and actual conscious mental states of a person *A* that has the neural system *N*, is equal to the number of all possible and actual Φ- structures that can be and are generated by that neural system *N*.

We can make this claim because there is one-to-one match between conscious mental states and their unique Φ-structures. Just like we can say that the number of all territories and the number of their unique, equally detailed maps are equal. Therefore, in terms of cardinality, both versions of the theory have something in common. That is:

(3) The number of all possible and actual conscious mental states of a person *A* that has the neural system *N*, is equal to the number of all

possible and actual MICSs/Φ-structures that can be and are generated by that neural system *N*.

The Argument for The Criterion

IIT's soundness has been an issue of dispute. Critics tried to formulate objections regarding various aspects of the theory. Some objections stirred up the allegedly problematic methodology of the theory concerning the bridge between the axioms and postulates. Some others issued so claimed absurd consequences of the theory (Aaronson, 2014). A community even labelled IIT as pseudoscience (Fleming et al., 2023). Despite the objections, IIT is still considered to be a worthwhile option by many neuroscientists and philosophers.

I will offer a criterion based on the consequences of the theory's identity claims. The criterion is focused on the number of possible conscious mental states of an average conscious agent and the number of MICSs/Φ-structures for that agent, with the intend to compare them. If these two numbers do not match, then the identity claim of the IIT will be shown to be false. This conditional is all there is to the criterion that I propose. Here are some definitions and argument for the criterion:

Definition 1.) *N* is the neural system of the agent *A*.[4]
Definition 2.) *B* is the class of all MICSs/Φ-structures that can be and are generated by *N*. Definition 3.) *C* is the class of all possible and actual conscious mental states of *A*. Definition 4.) *P* is class of all atomic propositions.
Remark: *C* contains conscious mental states from possible & actual worlds only in which
N is the one and only neural system of *A*.

[4] "*A*" stands for any arbitrary conscious agent, whose complexity regarding conscious mental states are average. "*N*" stands for the *complex*, (i.e., physical substrate which realizeses consciousness.) *N* is *A*'s neural system or *complex*, in the sense that consciousness generated by *N* is directly available for only *A*.

The argument:

(4) If IIT is true, then the cardinality of *B* is equal to the cardinality of *C*.
(5) Cardinality of *C* is not smaller than $\aleph_0$.
(6) Therefore, for IIT to be true, the cardinality of *B* must not be smaller than $\aleph_0$.

The conclusion (6) is the criterion that is said to be presented. It posits that cardinality of class
B not being smaller than $\aleph_0$, is a necessary condition for IIT to be true.

Justification for The Premises

The first premise follows from the identity claims and the previously mentioned entailment of the theory: (3). Since theory claims that there is an equality between the number of MICSs/Φ-structures and conscious mental states of a person, the class of all of the former, should not be smaller than the class of the all of the latter. This is a rather straightforwardly extractable claim, considering the previously made explanations for the identity claims of the theory.

However, the second premise requires further external support. It basically posits that the class of all possible and actual conscious mental states that an average person can experience, has the cardinality not less than $\aleph_0$. That is to say, there are at least infinitely many possible conscious mental states available for an ordinary person. This claim can be supported with reference to atomic propositions. The feature of being the most basic/simple structures subject to truth bearing, makes them a great candidate for multiplying possible conscious mental states. This simplicity regarding the content, enables ordinary conscious agents to think about them without any cognitive hardship or constraints. If we were talking about compound propositions, we would hesitate to say the same due to complexity of semantic content.[5]

[5] Large compound propositions are complex contentwise. One can plausibly say that some complex compound propositions constructed by infinite conjunctions or dijunctions, cannot be thought by any conscious agents, let alone ordinary average ones.

It seems intuitive that there is a possible world for every atomic proposition, in which they are consciously being thought by an ordinary conscious agent, say previously defined *A*. In other words, there is no atomic proposition which cannot be subject to conscious thought of *A*.[6]

This entails that there are at least as many possible conscious mental states for *A* as there are atomic propositions. Thus, the class *C* has the same or bigger cardinality than the class *P*, the class of atomic propositions that is. It is known that the class of all atomic propositions, *P*, has the cardinality of $\aleph_0$ in first order logic. Therefore, we can hold that *C*'s cardinality is not less than $\aleph_0$ (i.e., the second premise).

Another way to justify the second premise might involve with sensory originated conscious states: visuals, sounds, tastes, etc. We can say that there are infinitely many different possible scenes or compositions including various sensory elements that can be presented to the conscious agent *A*. with these set ups, *A* would consciously experience all these different sensory compositions. For example, one sensory composition might be: "*A* is eating yoghurt in a freezing winter day, on top of a building with traffic noises coming beneath." There would be a unique conscious mental state, experienced by *A* for this scene. It seems *prima facie* that we can change this composition with little minute changes such that we can create very big amount of different sensory compositions. In fact, we don't even have to stict to the normal everyday sensory inputs. Since, we are dealing with metaphysical possibilities, we can construct compositions with bizarre "alien" sensory inputs. One such bizarre composition might be: "*A* is hearing a metaphysically possible sound *D*, while being touched by a metaphysically possible object *T*, with the metaphysically possible surface texture *G*." In this composition, *A* would experience a completely different conscious experience than the normal everyday ones. Thus, the thought is: there are infinitely many sensory compositions of which, *A* can experience unique conscious experiences for. That amounts to saying that there are at least infinitely many conscious experiences available to *A*.

[6] I assume that the thesis of cognitive phenomenology is true (i.e., there are conscious phenomenal experiences for entertaining cognitive activities/propositional attitudes.

Objection From Cognitive Limits

A critic might try to question the second premise (5). The objection is mainly grounded in the cognitive limits of average persons. Due to the argument's heavy reliance on the average persons and their mental states, critics may hope to question what are the cognitive limits of such persons and whether those limits permit the argument to work? Here are two assumptions that a critic may hold to object to the argument:

(7) An average person *A* can only consciously think of/comprehend propositions that are below a finite threshold in terms of semantic content load.

Let's accept that this threshold is up to one thousand letters/characters, that is to say, an average person cannot consciously comprehend propositions whose natural language translations have more than one thousand letters. Actually, the precise number does not matter, as long as it is finite. The next assumption is more controversial:

(8) The majority/more than half of the propositions that an average person *A* can possibly consciously think of, can be uniquely translated into natural language (i.e., English).

The assumption basically states that most of the propositions that can be thought by an average conscious agent *A*, are translatable to English.

With (7) and (8), the critic might proceed with a permutation to show that the number of the majority of the propositions that an average person *A* can consciously think of is finite. Let's hold (8) and translate the majority of the propositions that can be consciously thought by average person *A*. In light of (7), we know that no translated proposition has more letters than one thousand. Then how many propositions can be translated? We know that the number cannot be bigger than 26^{1000}. Because, we know that there are only one thousand letter slots (the limit mentioned in (7)) and 26 total letters in English. The first slot can be filled with any of the 26 letters and any other slot as well. With permutation, we can calculate the total number of variations. The result would be: 26^{1000}. In fact, most of the strings produced would be meaningless bundles of letters.

With this calculation we can infer that the number of the majority of the propositions that an average human *A* can consciously think of, is smaller than a

finite number. If the number of the majority/more than half of these propositions is smaller than some finite number, then the number of all of them must also be smaller than a finite number. Therefore:

(9) Average person *A* can only consciously think of finite number of propositions.

This claim undermines the justification of the second premise (5), because in the justification, it is stated that every atomic proposition can be consciously thought by an average person *A*. Since, the number of atomic propositions is infinite, the justification for (5) and the claim (9) contradicts.

Possible Reply

It can be argued that the average person *A* with having the same neural system *N*, exists in some class of possible worlds, which has infinite cardinality. In some of them, *A* is a villager and in others, *A* lives with aliens due to abduction. All these different sceneries *A* is faced to, causes different phenomenal experiences within his mind. Let's say that each world in the class makes *A* experience a totally unique experience, such that there are no two worlds which are equal in terms of experiences they provide to *A*. Since there are infinite number of worlds in which *A* resides, there are infinitely many distinct possible phenomenal mental states of *A*. Let's try to depict each of these experiences with propositions such that for each experience there is a unique proposition:

(10) "I experience Γ right now."

"Γ" is a variable for every phenomenal conscious state that *A* can possibly experience. Thus, there are infinite versions of the proposition (10) in which *A* can possibly believe or think about. However, none of the versions could be written down due to the private, ineffable nature of phenomenal conscious experiences. Therefore, there are at least infinite number of propositions, which can be consciously thought by *A* and are untranslatable to English. That would mean that the conjunction of assumptions (7) and (8), namely the thesis that the number of conscious mental states of *A* is a finite number, is at least intuitively false.

Conclusion & Last Notes

The question whether IIT complies with the criterion seems to be answerable on empirical basis. However, I suspect that the number of MICSs or Φ-structures that can be *realized* by an average human is a very large finite number, yet finite nonetheless. The biological make-up of neural system seems not to be able to generate infinite number of different relevant mathematical structures. After all, MICSs and Φ-structures are mathematical forms that are *realized* by states of the neural system (i.e., *complexes*). An average person's neural system, however, cannot produce infinite number of different neural states/*complexes*. Therefore, the neural system of a human cannot be the *realizing* substrate of infinite number of mathematical structures (i.e., MICSs and Φ-structures). Due to these reasons, I foresee that the IIT would fail the criterion.

In this paper, I explained the identity claims of Integrated Information theory, along with the basic features of it. My intention was to provide a criterion or a necessary condition that the theory needs to fulfill for being true. I focused mainly on the identity claims of the theory, for they are the subject to the criterion that I propose. It is shown that both versions of the theory, 3.0 & 4.0, entail there to be an equality between the number of conscious mental states of a person, and the number of mathematical entities (i.e., MICS/Φ-structures) of that person's neural system. The criterion is dedicated to test whether these two numbers were equal. If they are shown not to be equal, then it would be concluded that the theory is false. In order to support the criterion, I presented an argument entertaining a comparison of cardinality between classes of relevant kinds. Then, I tried to justify the premises with relying on: cognitive phenomenology and metaphysically possible sensory compositions' unique conscious mental states. Later on, I considered an objection concerning the cognitive limits of average conscious agents and intends to object to the second premise (5) of the argument. I proposed a line of thought which utilizes possible phenomenal experiences of an average conscious agents, in order to demonstrate the falsity of the consequence of the objection.

References

Aaronson, S. (2014, May 21). *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. https://scottaaronson.blog/?p=1799

Albantakis, L., Barbosa, L. R., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., William, Alireza Zaeemzadeh, Mélanie Boly, Bjørn Erik Juel, Shuntaro Sasai, Fujii, K., Imhonopi, D., Hendren, J., Lang, J. P., & Giulio Tononi. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, *19*(10), e1011465–e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Fleming, S. M., Frith, C. D., Goodale, M. A., Lau, H., LeDoux, J. E., Lee, A., Michel, M., Owen, A. M., Megan, & Slagter, H. A. (2023). *The Integrated Information Theory of Consciousness as Pseudoscience*. https://doi.org/10.31234/osf.io/zsr78

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*(1), 4164. https://doi.org/10.4249/scholarpedia.4164

Tye, Michael, "Qualia", *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2021/entries/qualia/>.