The state of the Integrated Information Theory, its boundary cases and the question of ‘Phi-conscious’ AI

Tine Kolenik
Jožef Stefan Institute & Jožef Stefan
International Postgraduate School
Jamova cesta 39
1000 Ljubljana, Slovenia
+386 1 477 3807
tine.kolenik@ijs.si

Matjaž Gams
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
+386 1 477 3644
matjaz.gams@ijs.si

ABSTRACT

This work analyzes Giulio Tononi’s Integrated Information Theory of consciousness, defined in 2016, the tools it offers to calculate the level of consciousness in any given system, produced in 2018, and compares the theory to other relevant recent theories of consciousness. It then discusses issues with the theory as well as the tools, namely that they are unreliable due to a variety of shortcuts that give different approximations, as current technology does not allow faithful computation of consciousness, i.e. a system’s Phi. The testing confirms the problems with running time (*O*). Tononi’s stand on AI is then problematized in relation to IIT. The authors’ thoughts and treatise on a possibility of Phi-conscious AI is presented afterwards. AI systems are separated in three levels of hierarchy according to Marr and two types – knowledge representation-based and neural network systems according to Shoham. The authors hypothesize that combining both types brings AI closer to consciousness, which should hold true according to the multiple knowledge principle. Both systems are evaluated in relation to IIT’s axioms and postulates. Evaluation shows that their combination conforms to more axioms and postulates than both types do separately, therefore confirming the hypothesis. However, AI is still not Phi-conscious as it does not encompass all of IIT’s requirements.

Keywords

Artificial intelligence, consciousness, functionalism, Integrated Information Theory.

1. INTRODUCTION

Consciousness, this infinitely intimate state that we cannot escape and which encompasses our every thought, our every feeling and our every experience, is currently one of the most explored phenomena in science. It was explored with natural scientific methods more than 100 years ago by figures like the psychophysicists William James, Gustav Fechner, Hermann von Helmholtz and Wilhelm Wundt, but the research stopped as it was seen as a primitive, subjective and unscientific practice [1]. However, since the late 1990s, consciousness was again established as a phenomenon not only worth of exploring, but being able to be explored [2].

Theories of consciousness are abound, and there are many unique proposals, featuring orthogonal presuppositions, various ontological claims and sequestered methodologies for inquiry. Some of the most well received recent theories include the Global Workspace Theory [3], the Multiple Drafts Model [4], predictive coding approaches [5] and quantum theories of consciousness [6]. Among all, the Integrated Information Theory (IIT) of consciousness [7], proposed by the neuroscientist and psychiatrist Giulio Tononi, was described as the most formally sound, most computer science related and the most scientifically viable theory in this field yet [8].

IIT is based on a mathematical concept or quantity Φ, Phi, which can be calculated for any given system and represents *integrated information* (more in Section 3). IIT claims that integrated information is almost entirely correlated with the level of consciousness in the system Φ is calculated for. For example, the human brain has a very high Φ, which according to IIT, means that it is very highly conscious. But Φ can be calculated for any given system, so even atoms have some low number of Φ, or systems such as a light switch [9]. This conceptualization comes very close to the philosophical view of the mind called panpsychism, which proposes that consciousness or mind is a fundamental property of each and every part of any given system (from atoms to rocks to buildings to planets to the universe itself) [10]. This connection was also acknowledged by Tononi and Koch [11]. Another important aspect of IIT pertains to the hard problem of consciousness, which describes the explanatory gap between qualia or experience and physical states. IIT eschews the hard problem by presupposing consciousness as intrinsically real due to a system’s cause-effect powers upon itself (see Section 3, Axiom 1). This axiomatic property of IIT circumvents the hard problem debate, which is why it will also not be addressed any further in this work as it is out of its scope. The wider framework of IIT is described in Section 3. However, since even photodiodes’ Φ is above zero, the threshold for levels of semantically reasonable consciousness should be above zero in order to differentiate between what is commonly seen as conscious and unconscious. This should serve for easier discussions on consciousness in boundary cases such as artificial intelligence (AI).

In general, this paper is an upgrade of the paper by Gams [12], who presents an older version of IIT defined in 2014, offers a commentary on it and sets foundations for discussing AI in relation to IIT. The current work encompasses:

a. the state of the mentioned recent theories on consciousness in order to set them apart from IIT (Section 2),

b. an analysis of the state of IIT in its updated, newest form alongside with the recently developed tools available for measuring consciousness of any given system (Section 3), and

c. an analysis of the boundary cases for consciousness as described by Tononi [13] with the focus on AI and its possibilities for possessing consciousness (Section 4).

The paper ends with the authors' intentions for future work and some concluding thoughts.

2. STATE OF THE RECENT THEORIES OF CONSCIOUSNESS

This Section briefly presents the current state of the following theories on consciousness: the Global Workspace Theory [3], the Multiple Drafts Model [4], predictive coding approaches [5] and quantum theories of consciousness [6]. It also offers a short criticism of each and whether they encompass the possibility for AI to be conscious.

The Global Workspace Theory (GWT), which spawned many advanced off-shot theories such as the 'neuronal global workspace' theory [14], relies on the concept of global availability of conscious content. Conscious content is supposedly available to all cognitive processes (e.g., attention, decision-making), which are connected more to certain parts of the brain, while conscious content inhabits a global neuronal activity across the brain. Consciousness is therefore widely spread, while various processes and states compete for being brought into this conscious landscape. The theory can explain various neuronal phenomena as well as functional cognitive processes, but it is not clear on how the graduality (or binariness) of consciousness works and how to precisely measure it. If the organizational aspects of GWT were realized in computers, it would be sensible to say that computers would be conscious.

The Multiple Drafts Model is a cognitivist theory of consciousness and proposes that there is "no reality of conscious experience independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory)." [4, p. 132] The theory claims that there are numerous interpretations of the sensory data that comes in through our senses. Since these are processed in different parts of our brains at different times, the first of the multiple drafts that checks all the necessary boxes in the neural processing is the one that is acted upon, and that the experience accompanying it is illusory. However, critics claim that the theory does not hold the power to explain or predict neuropsychological research data. It also does not offer mathematical explanations. Regardless, Dennett believes that mental functions are functions in a mathematical sense, which means that they can be formalized in a machine, resulting in a conscious AI.

Predictive coding approaches [5] are probably the most recent approaches to understanding the mind. Predictive coding refers to the theory that the minds and brains are fundamentally prediction machines. The mind builds a hierarchical generative model of the world which it is always predicting. This radically changes the idea that the sensory input and information-processing of it is a feed-forward process, that sensory data travels from, e.g., the eye through the brain's multiple layers of processing, and in the end, causes a motor action. Instead, the brain predicts the next input to the eyes before the input appears. The theory is currently one of the most researched, if not the most researched theory in cognitive science [15]. Predictive coding is a highly mathematical theory, as it partly relies on computer science algorithms, meaning that it should be able to encode at least some aspects of what predictive coding has to say on consciousness in machines.

Quantum theories of consciousness mainly claim that classical mechanics cannot explain consciousness. It is quantum entanglement and superposition as well as other quantum phenomena that cause consciousness [6]. However, the quantum hypotheses mostly discuss how quantum phenomena may give rise to consciousness and not much about the consciousness itself. The main (and particularly enormous) problem is that they are nowhere near testable. Since the quantum theories rely on quantum phenomena in terms of consciousness existing, machines first need to possess these quantum phenomena. Then, according to the theory, they can be built to have consciousness.

This collection of various contemporary theories of consciousness tries to sketch the state of consciousness theories so that IIT is placed in context and that it can be evaluated against them. The next Section discusses the state of IIT.

3. STATE OF THE INTEGRATED INFORMATION THEORY

This Section more thoroughly introduces IIT and the recently released tools and methods for measuring Φ. This serves as a continuation and an upgrade of the description of IIT by Gams [12] as well as a foundation on which Section 4 analyzes AI in regards to Φ.

The IIT takes inspiration from various sources – panpsychism was already mentioned – but it starts from getting away from purely searching for neuronal and behavioral correlates of consciousness and experience. It asks the harder questions of why cerebral cortex gives rise to consciousness but not cerebellum, even though it has approximately 4 times more neurons than the cerebral cortex and of what is important for consciousness in terms of various boundary cases having it. The latter is especially important, and Tononi and Koch [11] list a number of such cases where they ask whether they are conscious or not: 1) patients and infants, 2) animals, and 3) machines (more on this in Section 4). IIT therefore does not want to only work with collected data on cases where consciousness is freely attributed – neurotypical adult humans – it wants to propose what consciousness and experience are and what kind of systems in regards to their interactional properties can have them. IIT does that, however, in a reverse order than what consciousness researcher usually do – it starts from experience by positing five axioms and deriving five postulates that describe systems for which the axioms are true. On top of that, IIT establishes a calculus for precise measurements of consciousness, which it connects to integrated information, symbolized by Φ, Phi.

The five axioms and postulates are:

1. Intrinsic experience:

Axiom: Consciousness is real, and it is real from its own perspective.

Postulate: System must have cause-effect power upon itself.

2. Composition:

Axiom: Consciousness is composed of phenomenological distinctions, which exist within it.

Postulate: System must be composed of elements that have cause-effect power upon the system.

3. Information:

Axiom: Consciousness and each experience is specific, differing from other possible experiences.

Postulate: System must possess cause-effect sets that differ from each other in their space of possibilities.

4. Integration:

Axiom: Consciousness is unified and experience is irreducible to a set of its phenomenological distinctions taken apart.

Postulate: System must specify its cause-effect structure as to be unified, irreducible to mere sum of its parts ($\Phi_{\text{system}} > \Phi_{\text{sum of parts}}$).

5. Exclusion:

Axiom: Consciousness and experiences are definite and are the way they are, nothing else.

Postulate: System must specify its cause-effect structure to be definite, always over a single set of elements and maximally irreducible ($\Phi_{\text{system}} > \Phi_{\text{any given sub-system}}$).

The remaining part of this Section focuses on the notion of integrated information, Φ, as this is the part of IIT that Tononi's team is paying attention to the most in the recent years in terms of updating and revising it, especially with new tools.

Among others, the notion of integrated information offers the answer to the question of why cerebral cortex generates consciousness, but not cerebellum, even though the later has four times more neurons than the first. It also explains how even photodiodes can have experience and therefore, albeit very low level of, consciousness.

The main idea behind Φ and why it measures consciousness is this: First, it measures information in a certain system. This information is denoted by how much information the system has about itself, which is defined as a number of possible states, past and future. Second, this measure of information is coupled with how this information is integrated. What is measured is how much the information depends on the interconnectedness of the system's parts. To demonstrate this measurement, the system is split (into an arbitrary number of sub-systems) and then information is measured again. The more information that is lost, meaning the more information that arose from this interconnectedness, the more integrated the system was. Integration is also the reason why Tononi argues that computers have very little consciousness – because even though they can have much information, it is not integrated. He argues that transistors (he deems the physical, implementational level the most important) do not lose much structure or information if split, as they can still give rise to the same system (more on this in the next Section).

However, measuring Φ, even if we generally know what we want to measure, is extremely difficult. The biggest problem is that Φ cannot be calculated with our current computational technologies even if the system is only as big as a few nodes. Φ can be approximated with various different shortcuts and heuristics, but the problem is that for the same system, the approximation wildly varies depending on the technique for the approximation used [16]. In 2018, Mayner et al. [17] produced PyPhi, a Python software library that allows one to study the cause-effect structure of a given system in relation to IIT and calculate Φ. However, even though it encompasses a number of heuristics to calculating Φ, the algorithm's running time is exponential in terms of number of nodes increasing. Currently, the algorithm's running time is $O(n53^n)$, where n denotes the number of nodes. Running simple CPU experiments, it takes 24 hours to calculate Φ using the major complex of systems approach on a seven-node system if run on 4 × 3.1GHz CPU cores (see Table 1). Other shortcuts produce different running times, but also different Phis.

Table 1: Test of running time of Φ calculations for three systems with a different number of nodes.

# of nodes in system	Running time
3	~8 seconds
5	~2.5 minutes
7	~24 hours

The running time and the problem of getting different Phis with different calculations is one of the biggest criticisms of IIT. It also seems that in its current version, V3, IIT does not provide falsifiable predictions, which is one of the most common criticisms of most theories of consciousness.

4. INTEGRATED INFORMATION THEORY AND ARTIFICIAL INTELLIGENCE

This Section speculates on conscious AI in relation to IIT, dubbed as Phi-conscious AI. The authors address some of Tononi's points on AI, argue that some of his points may not be correct regarding it, propose that AI on certain levels may be seen as conscious and evaluate different AI paradigms through IIT's axioms.

Tononi examines AI only from a physical level. He only considers what computers are physically made of and makes claims exclusively about transistors and their inability to reach high Φ due to not being integrated – if one splits transistors, they can still possess the same information value. Tononi even states that if "integrated information theory is correct, computers could behave exactly like you and me, and yet there would literally be nobody there" [18, para. 32]. This means that even if they were programmed to satisfy the axioms and have a sufficiently high Φ, according to Tononi, their physical, transistor-based implementation would preclude 'true' consciousness. AI that would behave perfectly humanly would be the philosophical zombie. However, Tononi takes a very narrow perspective on AI that may even be in contention with IIT itself, as IIT's axioms and postulates do not necessarily require the implementational level of a system to be the one that counts in term of consciousness. Marr [19] proposes a three-level hierarchy in regards to AI and cognition in general: 1) computational level (what the system does and why), 2) algorithmic level (how the system does what it does), 3) physical level (the realization of the first two levels). The first two levels may bear a much higher Φ. However, the computational level does presuppose some functionalist ideas, namely that mental states are as they are because of the function they perform.

To speculate on whether certain types of AI on the 1st and 2nd level of Marr's hierarchy are Phi-conscious, AI is separated in three categories. It is investigated whether IIT's axioms and postulates hold true for them. The AI categorization is based on Yoav Shoham's invited talk [20] at this year's International Joint Conferences on Artificial Intelligence (IJCAI), one of the biggest and oldest AI conferences in the world. Shoham categorizes AI in roughly two categories: knowledge representation (KR) based AI (commonly dubbed as 'good old-fashioned AI') and neural networks (NN). His hypothesis is that KR is good for certain problems, that NN is good for other problems and that by combining the two, AI will enter a new era of progress as KR+NN will work better than its parts (see Figure 1). Our hypothesis mirrors Shoham's – we believe that KR may satisfy some IIT's axioms and postulates, that NN may satisfy some other axioms and postulates, but that together they would have higher Φ than they would if treated separately and then summed up. This thinking is also based on the multiple knowledge principle [21], according to which our hypothesis should hold true.

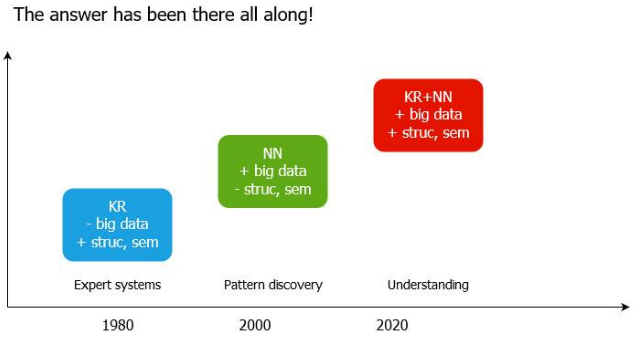

Figure 1: Shoham's vision for AI (struc = structured, sem = semantics). Adapted from [20].

KR mostly encompasses expert systems. These are systems that have all their domain knowledge programmed into them with various rules, which are explainable and symbolic in nature. The process of knowledge acquisition is top-down, meaning that the designer presupposes everything they know.

NN encompass learning systems that usually look for patterns. Their knowledge is produced from lots of data (big data), bottom-up, they are subsymbolic and very robust.

The table below (Table 2) shows the analysis for KR, NN and KR+NN in relation to IIT's axioms and postulates. KR+NN's relation to IIT is determined by using logical disjunction (∨, (x)or) between KR and NN, as axioms and postulates have to hold true only for one to hold true for KR+NN.

The arguments in Table 2 claim that by combining KR and NN, AI gets closer to achieving consciousness according to IIT. What seems to be lacking in both is *exclusion*. AI therefore cannot be characterized as being Phi-conscious just yet, but our initial hypothesis is confirmed.

There is more to IIT's problems regarding AI. One problem is that Tononi clearly states that his theory should be judged according to how it explains the empirical data about consciousness [11]. There is a problem with this in relation to AI – there is no empirical data about consciousness. Tononi presupposes consciousness and acts accordingly – that neurological data on the brain is in fact empirical data about consciousness, without calculating Φ to find out whether this is true. This inherently cripples meaningful research on AI consciousness, as one cannot do the same and presuppose it in, e.g., robots. You cannot, as Tononi tries to do with IIT, reverse engineer the process of scientific investigation and theorizing.

Table 2: Analysis of KR, NN and KR+NN in relation to IIT's axioms and postulates.

AI type IIT	KR	NN	KR+NN (KR ∨ NN)
Intrinsic experience	can have cause-effect power upon itself, as rule-based system may operate on feedback loops and recursions (the specified rules may change) that are being performed without input **TRUE**	layers may easily be interconnected or connect in a way (bi-directional layers, feedback loops on the same nodes …) for NN to have cause-effect power upon itself, especially in no-input NNs such as (generative NN, Kohonen NN …) **TRUE**	**TRUE**
Composition	has strong compositional property; computational rules may be linked between each other and have effect among each other **TRUE**	due to the self-organizational nature of NNs, modularity and therefore composition is not clear and entirely explainable; nodes do connect, but may not hold true for concepts; since it is very robust, parts may be removed without affecting the system itself **FALSE**	**TRUE**
Information	can possess many cause-effect sets, differing from each other (Tononi also states that machines have high information value) **TRUE**	a number of cause-effect sets is usually operationally the same in relation to their power in the system (which is why optimization by reducing NN size works) **FALSE**	**TRUE**
Integration	in KR, the sum of its parts by definitions cannot be more than the system itself, as expert systems are inherently modular, therefore violating '$\Phi_{\text{system}} > \Phi_{\text{sum of parts}}$' **FALSE**	works as a unified and distributed system and completely irreducible to the sum of its parts as nodes necessarily organize between each other in an inseparable way; '$\Phi_{\text{system}} > \Phi_{\text{sum of parts}}$' holds true **TRUE**	**TRUE**
Exclusion	cannot guarantee that a KR system is a maximally irreducible, especially due to its modularity, therefore violating '$\Phi_{\text{system}} > \Phi_{\text{any given sub-system}}$' **FALSE**	Usually a NN can be reduced to an operationally equally effective subsystem that has the same integration and information values (which is why optimization by reducing NN size works), which implies that NN systems violate '$\Phi_{\text{system}} > \Phi_{\text{any given sub-system}}$' **FALSE**	**FALSE**

5. CONCLUSIONS AND FUTURE WORK

This work presents the latest iteration of the Integrated Information theory proposed by Tononi, some tools the IIT researchers offer for calculating Φ, and the problems of both. Some other theories of consciousness are presented as well to put IIT in context, especially in regards to AI. The biggest contribution of this work is in trying to speculate on whether AI is, as dubbed by the authors, Phi-conscious or not. We speculate about consciousness on various types of AI, categorized by Shoham, and hypothesize that combining different types brings us closer to Phi-conscious AI, which we claim to confirm (Table 2).

Our future work includes more thorough analysis of different concrete KR and NN systems, but our foremost interest lies in working with KR+NN systems. This seems to be the future regardless of IIT, but we want to make KR-NN systems as close to Phi-conscious as possible and see what consequences will emerge. Other ideas for future work include using machine learning and state-of-the-art algorithms to deal with the algorithm running time better in terms of developing heuristics to shorten the calculating time, and consequently calculating Phi for systems such as recurrently connected Turing machines to find out whether it is higher than the sum of individual Turing machines due to dynamic interactions [21].

6. REFERENCES

[1] Hawkins, S. L. 2011. William James, Gustav Fechner, and Early Psychophysics. *Front. Physiol*. 2, 68 (2011). DOI= 10.3389/fphys.2011.00068.

[2] Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Oxford.

[3] Robinson, R. 2009. Exploring the "Global Workspace" of Consciousness. *PLoS Biol*. 7, 3 (2009), e1000066. DOI= 10.1371/journal.pbio.1000066.

[4] Dennett, D. C. 1991. *Consciousness Explained*. Little, Brown & Co, Boston, MA.

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

[6] Atmanspacher, H. 2019. Quantum Approaches to Consciousness. *The Stanford Encyclopedia of Philosophy*, (2019).

[7] Tononi, G., Boly, M., Massimini, M., and Koch, C. 2016. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci*. 17, 7 (2016), 450-461. DOI= 10.1038/nrn.2016.44.

[8] Gennaro, R. J. 2018. *Consciousness*. Springer, Cham.

[9] Tononi, G. 2008. Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Bulletin*. 215, 3 (2008), 216-242. DOI= 10.2307/25470707.

[10] Alter, T. and Nagasawa, Y. 2015. *Consciousness in the Physical World: Perspectives on Russellian Monism*. Oxford University Press, Oxford.

[11] Tononi, G. and Koch, C. 2015. Consciousness: here, there and everywhere? *Phil. Trans. R. Soc. B*. 370 (2015). DOI= 10.1098/rstb.2014.0167.

[12] Gams, M. 2015. Kochove meritve zavesti. In *Cognitive science: proceedings of the 18th International Multiconference Information Society – IS 2015* (Ljubljana, Slovenia, October 8-9, 2015). Jožef Stefan Institute, 11-14.

[13] Oizumi, M., Albantakis, L, and Tononi, G. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput. Biol*. 10, 5 (2014), e1003588. DOI= 10.1371/journal.pcbi.1003588.

[14] Dehaene, S. 2015. *Consciousness and the Brain*. Viking, New York.

[15] Bohannon, J. 2018. *A computer program just ranked the most influential brain scientists of the modern era*. Retrieved September 11, 2019, from https://www.sciencemag.org/news/2016/11/computer-program-just-ranked-most-influential-brain-scientists-modern-era.

[16] Bayne, T. 2018. On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 2018, 1 (2018), niy007. DOI= 10.1093/nc/niy007.

[17] Mayner, W. G. P., Marshall, W., Albantakis, L, Findlay, G., Marchman, R., and Tononi, G. 2018. PyPhi: A toolbox for integrated information theory. *PLoS Comput. Biol*. 14, 7 (2018), e1006343. DOI= 10.1371/journal.pcbi.1006343.

[18] Robson, D. 2019. Are we close to solving the puzzle of consciousness? Retrieved September 11, 2019, from http://www.bbc.com/future/story/20190326-are-we-close-to-solving-the-puzzle-of-consciousness.

[19] Marr, D. 2010. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, Cambridge, MA.

[20] Shoham, Y. *Award for Research Excellence presentation. Invited Lecture*. 2019 International Joint Conference on Artificial Intelligence.

[21] Gams, M. 2001. *Weak intelligence: through the principle and paradox of multiple knowledge*. Nova Science, Hauppauge, NY.