Linguistic and Philosophical Investigations
ISSN: 1841-2394, e-ISSN: 2471-0881
Vol 23 (1), 2024
pp. 1655 - 1662

Integrated Information Theory and Panpsychism

[1]Carlos Blanco

[1]Universidad Pontificia Comillas (Madrid, Spain); cbperez@comillas.edu

Abstract: Integrated Information Theory (IIT) has been challenged, among other reasons, for its supposed commitment to panpsychism (the belief that mental qualities pertain, in different degrees, to all forms of being, not just to entities endowed with highly complex brains). Here I want to argue that in order to avoid accusations of this kind IIT must revisit the internal consistency and explanatory completeness of its axioms, by clarifying the set of conceptual restrictions that may overcome potential panpsychistic commitments.

Keywords: Integrated Information Theory, panpsychism, consciousness, axioms, restrictions.

Received: 07 May 2024 **Revised:** 13 June 2024 **Accepted:** 24 June 2024

1. Integrated Information Theory

Integrated Information Theory (IIT) appears as one of the most promising models for explaining consciousness in scientific terms (Tononi et alii, 2016). Together with its main rival, the Global Workspace Theory (Dehaene-Naccache, 2001; Baars, 2005; Mashour et alii, 2020), it aims to provide a testable conceptual framework for addressing the mystery of the nature and properties of consciousness (for a review of the main contemporary theories of consciousness, see Yaron et alii, 2022; Seth – Bayne, 2022). The ultimate goal is to attack the problem of subjective experience objectively, in order to explain the causal connection between physical and psychological states, answering the question of how a particular subjective experience is generated by its neural correlates.

If Global Workspace Theory is attractive for its proximity to neuroscientific knowledge, in particular to our understanding of large-range neural connectivity, IIT offers the possibility of applying concepts that have revealed their fruitfulness in other domains of scientific and mathematical inquiry, like those of information and integration.

Nevertheless, recently the scientific status of IIT has been challenged by several prominent researchers in the fields of psychology and consciousness studies. Accusations cover a wide range of topics, and have led to claims about its pseudoscientific character. In a letter published in *PsyArxiv*, these authors state that "according to IIT, an inactive grid of connected logic gates that are not performing any useful computation can be conscious—possibly even more so than humans; organoids created out of petri-dishes, as well as human fetuses at very early stages of development, are likely conscious according to the theory; on some interpretations, even plants may be conscious. These claims have been widely considered untestable, unscientific, 'magicalist', or a 'departure from science as we know it'. Given its panpsychist commitments, until the theory as a whole—not just some hand-picked auxiliary components trivially shared by many others or already known to be true—is empirically testable, we feel that the pseudoscience label should

indeed apply. Regrettably, given the recent events and heightened public interest, it has become especially necessary to rectify this matter." (Fleming et alii, 2023).

Thus, it seems clear that panpsychism is one of the key targets of IIT's detractors. More specifically, the idea that consciousness would be potentially ubiquitous throughout nature, as manifested in Aaronson's inactive grid of connected logical gates (which, in conformity with IIT, might be endowed with some degree of consciousness, and therefore with *phi*>0; see Horgan, 2015), a possibility that implies a sort of *reduction ad absurdum* for the theory. I will leave aside the accusation of pseudoscience. Is any untested hypothesis pseudoscientific? Is string theory pseudoscience because it makes no testable predictions within the present scope of technology? Was relativity pseudoscience until it was tested by observation and experiment? Rather than this, should the term pseudoscience not be reserved to hypotheses that, by their formulation, defy any way of empirical testing, or make predictions outside any conceivable range of validation?

A discussion on the ontological basis of IIT is therefore pertinent. Is IIT a modern version of philosophical panpsychism? Can it be associated with a different metaphysical school? What ontological tenets does it contain, and what picture of reality is suggested by its axioms and postulates?

Some authors have stressed the metaphysical affinity between IIT and funcionalist emergentism, based on the conceptual connection between integration and emergence (Negro, 2022). In some versions of this emergentist interpretation, "phenomenal consciousness is a functional emergent property of integrated systems, in a strong ontological sense" (Cea, 2020: 2200), and subjective qualities would be considered as global-level emergent functional-realizers. Yet, the relationship between IIT and some forms of panpsychism has been highlighted by different authors (Morch, 2019; Owen, 2019; Sánchez-Cañizares, 2022), and suspicions concerning its panpsychist commitments cannot be eluded easily.

According to panpsychism, mind is as fundamental as matter (Bruntrup-Jaskolla, 2016). Unlike physicalism, it does not try to reduce mind to matter, because it considers both mind and matter as equals in ontological status. Mental properties would therefore pertain to any form of reality; the ultimate form of what there is would be neither mind nor matter, since both are equally fundamental, perhaps manifestations of a deeper structure of reality. Here, one can find reminiscences of neutral monism, even of dualism (at least if one adheres to a strong distinction between mind and matter, rather than a merely descriptive separation between both ontological states). Although a discussion on the connection between both approaches to the nature of mind and matter transcends the scope of this letter (for a detailed analysis, see Holman, 2008), it may suffice to state that there is, certainly, a profound conceptual similarity between the two paradigms, panpsychism and neutral monism, the main difference residing, perhaps, in the tendency of the first to underline the ontological primacy of consciousness over matter, rather than simply affirming that they are two sides of the same coin.

Indeed, a theory that attributes —at least potentially— mental properties, in particular the possibility of having conscious experiences, to all forms of matter seems not only at odds with empirical evidence but even with conceptual consistency, as it does not offer a proper way of differentiating consciousness from other natural phenomena.

2. The axioms: consistency and completeness

To what extent is IIT committed to panpsychism? By reviewing the principal axioms of this theory, derived from the phenomenology of consciousness, and the postulates inferred from the axioms, it is possible to discern a set of conceptual restrictions that invalidate the general claim about its intrinsic connection with panspychism. Thus, although it is true that a pansychist interpretation of IIT is legitimate, it is by no means necessary.

Such axioms, meant to capture the essential properties of consciousness in a complete and consistent way, are taken as both evident (that is, as immediately given to any observer) and mutually independent. They "state that every experience exists intrinsically and is structured, specific, unitary and definite. IIT then postulates that, for each essential property of experience, there must be a corresponding causal property of the PSC [the physical substrate of consciousness]. The postulates of IIT state that the PSC must have intrinsic cause–effect power; its parts must also have cause–effect power within the PSC and they must specify a cause–effect structure that is specific, unitary and definite." (Tononi et alii, 2016: 450). Furthermore, "based on the postulates, it permits in principle to derive, for any particular system of elements in a state, whether it has consciousness, how much, and which particular experience it is having. IIT offers a parsimonious explanation for empirical evidence, makes testable predictions, and permits inferences and extrapolations." (Tononi, 2015: 5).

The axioms are the following:

Intrinsic existence: this axiom states that consciousness exists; that it is a real phenomenon, immediately given to my intuition (therefore, it is reminiscent of Descartes' *cogito*). Consciousness is primary in the sense that one can be absolutely sure of its existence through internal analysis (i.e., reflection). Even if one can have doubts concerning the reality of external phenomena, one cannot be skeptical about consciousness, as it is immediately manifested as soon as one thinks.

Composition: conscious experience is composed of multiple "sub-experiences", and therefore of multiple, differentiated objects and qualities that can be thought of as conforming that particular conscious experience.

Information: each conscious experience is specific, as it offers a particular piece of information. It refers to specific objects or qualities, *differentiated* from other objects or qualities. This specific character underlies its ability to be informative.

Integration: this axiom is complementary to the axiom of information. If information stresses differentiation, and therefore the specific character of that conscious experience, integration addresses the fact that consciousness is unified: even if it is composed of different objects and qualities, it nonetheless constitutes a unitary experience, irreducible to its elements. Conscious experience can be considered a whole that transcends the sum of its parts. Clearly, this axiom encompasses the content of what is generally known as "the binding problem": the totality of my experience is perceived as a unitary phenomenon, even if it is still possible to detect its constituent parts. This axiom allows us to "measure" consciousness, "calculated as the distance between the conceptual structure specified by the intact system

and that specified by its minimum information partition." (Tononi et alii, 2016: 452).

Exclusion: according to this axiom, consciousness is definite in spatio-temporal terms.

In synthesis, "the information axiom asserts that every experience is specific – it is what it is by differing in its particular way from a large repertoire of alternatives. The integration axiom asserts that each experience is unified – it cannot be reduced to independent components. The exclusion axiom asserts that every experience is definite – it is limited to particular things and not others and flows at a particular speed and resolution. IIT formalizes these intuitions with postulates." (Tononi, 2012: 290). Such postulates are logical derivations from the axioms, and pave the way for an objective, third-person perspective on conscious processing in different systems.

Now, the question comes regarding the physical substrate of consciousness (PSC). A purely formal exposition of the theory may give the impression that such physical substrate is not necessarily attached to a neural correlate. Logical gates of different kinds, as far as they admit the reception of inputs and outputs, could in principle satisfy the axioms of IIT. Because this theory attempts at developing a mathematical model to evaluate the quality and quantity of conscious experience, a critic might argue that its "top-down" approach (rather than a "phenomenology-first", because of its axiomatic, deductive nature —even if these axioms have been attained by induction from our individual phenomenological experience—; for a critique of this axiomatic methodology, see Bayne, 2018: 2-3), in which one starts with a set of axioms obtained by pure phenomenological observation of the properties of consciousness, instead of inducing the features of consciousness from the analysis of cerebral processes, leaves room to attributing consciousness to systems other than a highly developed brain. Indeed, IIT "does not start from the brain and ask how it could give rise to experience; instead, it starts from the essential phenomenal properties of experience, or axioms, and infers postulates about the characteristics that are required of its physical substrate." (Tononi et alii, 2016: 450).

In other words, the theoretical independence between the formal apparatus of the theory, in terms of axioms susceptible to mathematical expression, and the specific neural structures that should constitute its physical substrate might mean that the hypothetical causal mechanism responsible for generating a conscious experience is not attached to a neural correlate in the cortico-thalamic system, as assumed by a majority of models about the nature and properties of consciousness.

Yet, this assertion seems in contradiction with the efforts of IIT's supporters to link the postulates (derived from the axioms) and the physical substrate in terms of neural correlates (NCC). Tononi has written that "specifically, it states that the content-specific NCC correspond to the neural elements of the PSC in a particular state (activity pattern), which specify a particular phenomenal content; the full NCC correspond to the neural elements constituting the PSC irrespective of their particular state; the background conditions are factors that enable consciousness." (Tononi et alii, 2016: 452).

Furthermore, in light of the mathematical formalism used by IIT it is clear that there is a great number of entities with *phi*=0, which are therefore unconscious in accordance with the predictions of the theory. Indeed, the simplest conscious system contemplated by IIT (which would have *phi*>0) turns out to be a

feedback dyad (a system of two particles in a feedback loop [Doering et alii, 2019; McQueen – Tsuchiya, 2023)]).

3. Conceptual restrictions and the elasticity of the theory

From a conceptual point of view, at first glance the axioms of IIT seem so general in nature that it is legitimate to underline the logical separation between the properties of consciousness and their strict dependence upon a specific physical system, like the human brain. The qualities of intrinsic existence, composition, information, integration and exclusion could belong to a different physical substrate.

Two observations. First, rather than being a limitation for the theory, its degree of flexibility could be one of its greatest virtues. Its substrate-neutral approach enables the attribution of consciousness to artificial systems, leaving the discussion concerning machine consciousness open. Second, through the assumption of these axioms IIT is not committing itself necessarily to panpsychism. On the contrary, it is simply stating that any system capable of supporting a phenomenon in possession of these features could be considered conscious. It is left to empirical evidence whether such a system exists in the real world beyond highly complex brains. Indeed, the postulates derived from the axioms are notably restrictive in terms of determining what can be conscious and what cannot. The requirements posed by the postulates are so hard to fulfil that it is difficult to think of any other system susceptible to satisfying them. At least, and in principle, such hypothetical physical system should be able to create internal representations of the world, because the sum of the axioms (in particular, the combination of the axioms of composition and information) entails that the agent having conscious experiences is capable of "detaching" itself from the object, which falls under its power of analysis. The very idea of experience points in this direction. The question would be the following: is there any conceivable experience if the agent is unable to separate itself from the object that elicits its mental experience (and therefore its internal representation; see Blanco, 2020: 359ff.)? Whether or not a merely behavioristic account could suffice in some cases, if one admits that consciousness has intrinsic existence, rather than being epiphenomenal, the assumption of an internal realm in charge of forming representations of the external world seems inevitable.

Once more, apart, perhaps, from certain artificial systems of the kind that have been built recently in the domain of generative artificial intelligence, what natural systems can satisfy these axioms? What other systems in the world are known to be able to manage information in so complex ways? Instead of diluting consciousness by predicating it of any natural system, one may claim that IIT is emphasizing the specificity and complexity of this experience.

The only exception would perhaps lie in some sort of "transition systems" between the non-conscious (where *phi*=0) and the minimally conscious (with *phi*>0, yet to a minimal degree, such that a local maximum may be attained), like Aaronson's inactive grid of connected logic gates, McQueen's feedback dyad, and Oizumi's photodiode (Oizumi et alii, 2014), to which some minimal level of consciousness might be attributed, based on the postulates of IIT. In these cases, there is certainly an intriguing proximity to predicating consciousness of systems so simple that they would lack mental states. Nevertheless, in my opinion these elementary systems should not be labeled as conscious, even if their *phi* is greater than 0. They may possess a minimal degree of causal integration of information, manifested in *phi*>0, yet it is in

contradiction with the conceptual restrictions posed by the axioms and postulates of IIT to attribute consciousness to these objects.

In any case, this apparent coexistence of continuity and discreteness is one of the deepest conceptual and empirical objections to many models of consciousness. Indeed, the question is whether any theory of consciousness could liberate itself entirely from assuming some sort of "spectrum" of systems that, mediating between the unconscious and the fully conscious (if anyone knows what this would mean), possess a somehow vague ontological status. Perhaps, these systems might distinguish themselves by thresholds, each of which would constitute an "all or nothing" departure from the previous one.

The problem would therefore consist in clarifying to which value of *phi* corresponds the true frontier between the unconscious, the minimally conscious, and the fully conscious, given that it might be reasonable, in agreement with the theory, to admit the presence of a transition phase between complete absence of consciousness and a minimal degree of conscious experience. This is, indeed, what IIT has done (Tononi et alii, 2016), by differentiating the values of *phi* between minimally conscious states and wakefulness. Yet, the critics may still pose the question regarding what a large value of *phi* would look like. Should the scale be established on merely *a posteriori* grounds, by observing natural systems and comparing them to the predicted values of *phi*, or a theory as ambitious and encompassing as IIT should be capable of predicting from its axioms and postulates a convincing table of values? How many free parameters would be left in this process, that should be calculated *ex post*?

This ambiguity may open the window for contemplating, scientifically and philosophically, how consciousness can be conceived even without mental states and internal representations; something that defies many elements of the philosophical imagination, and seems contradictory with our previous statements. After all, a complex brain, in which mental states are possible, might not be the only system capable of satisfying the postulates of IIT. Yet, one must recognize that this is one of the principal difficulties of IIT —and perhaps of any theory of consciousness at all—. It might lead some sceptics to challenge its explanatory viability, given that it establishes a set of restrictions on the one hand, while leaving the door open for violating them on the other hand. Is this ambiguity calculated? Does it contribute to encapsulating the complex and evanescent nature of consciousness?

In any case, and as a final remark, I think that the key element of judgment here affects the axiom of intrinsic existence. In my view, IIT does not explain why should consciousness as such exist (indeed, it takes its existence for granted and elevates it to the category of an axiom), if the local maximization of causal power described by this theory perhaps could be attained without a concomitant conscious experience. Likewise, "any recurrent network can be unfolded into a feedforward network implementing the same function. In particular, any behavioural experiment can be seen as an input-output function, and can thus be implemented by both recurrent and feedforward networks." (Doerig et alii, 2019: 52). Thus, if the same function can be realized in different ways, given that feedforward and recurrent networks can be functionally equivalent, what is the real predictive value of *phi* and the true explanatory power of IIT?

Nevertheless, these are different questions, that transcend the scope of the article. My goal has not been to challenge IIT as such but to show that accusations of panpsychism are unfair, and may be refuted. Yet, I have doubts about the ability of this model to offer an explanation of consciousness that is theoretically consistent and empirically testable. I think that IIT measures the degree of causal integration (McQueen – Tsuchiya, 2023:3), which may be a necessary condition for consciousness, but it unlikely constitutes a sufficient condition for conscious experience (indeed, we do not know whether this property is susceptible to quantification). However, be it right or wrong regarding the ultimate nature of consciousness, what seems clear to me is that it is not necessarily committed to panpsychism.

4. Conclusion

A purely conceptual analysis of IIT's axioms (which, having been inferred from the phenomenology of consciousness, are not the result of scientific induction from observation and experiment) and the derived postulates shows that its commitment to philosophical panspychism should not be accepted so easily.

While its axiomatic formulation, somehow detached from specific neural correlates, leaves the window open for the question of machine consciousness (and possibly of other physical systems), it is clear that the postulates summarizing the main properties of conscious experience do not imply, necessarily, the assumption of panpsychism. On the contrary, they impose a set of restrictions regarding the class of physical systems that may satisfy their conceptual requirements.

Thus, rather than being too lax, by analyzing their content one may argue that they constrain the number of physical systems susceptible to fulfilling them. However, this view demands revisiting the consistency and completeness of the axioms and admitting that not all systems with *phi*>0 are automatically conscious.

References

[1] Baars, B. J. (2005). "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience." *Progress in brain research*, *150*, 45-53.

[2] Bayne, T. (2018). "On the axiomatic foundations of the integrated information theory of consciousness." *Neuroscience of consciousness* 2018, 4/1, 1-8.

[3] Blanco, C. (2020). *The integration of knowledge*. Peter Lang, New York.

[4] Bruntrup, G., & Jaskolla, L. (eds.). (2016). *Panpsychism: contemporary perspectives*. Oxford, University Press.

[5] Cea, I. (2021). "Integrated information theory of consciousness is a functionalist emergentism." *Synthese* 199.1-2, 2199-2224.

[6] Dehaene, S., & Naccache, L. (2001). "Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework." *Cognition*, *79*(1-2), 1-37.

[7] Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). "The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness." *Consciousness and cognition* 72: 49-59.

[8] Fleming, S. M., Frith, C., Goodale, M., Lau, H., LeDoux, J. E., Lee, A. L. F., … Slagter, H. A. (2023, September 16). "The Integrated Information Theory of Consciousness as Pseudoscience." https://doi.org/10.31234/osf.io/zsr78.

[9] Holman, E. (2008). “Panpsychism, physicalism, neutral monism and the Russellian theory of mind.” *Journal of Consciousness Studies* 15.5: 48-67.

[10] Horgan, J. (2015). “Can Integrated Information Theory Explain Consciousness?”. *Scientific American*. http://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/

[11] Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). “Conscious processing and the global neuronal workspace hypothesis.” *Neuron*, *105*(5), 776-798.

[12] McQueen, K. J. - Tsuchiya, N. (2023). “When do parts form wholes? Integrated information as the restriction on mereological composition.” *Neuroscience of Consciousness*, 2023/1, 1-11.

[13] Mørch, H. H. (2019). “Is the integrated information theory of consciousness compatible with russellian panpsychism?”. *Erkenntnis*, *84*(5), 1065-1085.

[14] Negro, N. (2022). “Emergentist Integrated Information Theory.” *Erkenntnis*, 1-23.

[15] Oizumi, M. - Albantakis, L. -Tononi, G. (2014). “From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0.” *PLoS computational biology*, *10*(5), e1003588.

[16] Owen, M. (2019). “Exploring common ground between integrated information theory and Aristotelian metaphysics.” *Journal of Consciousness Studies*, *26*(1-2), 163-187.

[17] Sánchez-Cañizares, J. (2022). “Integrated information theory as testing ground for causation: Why nested hylomorphism overcomes physicalism and panpsychism.” *Journal of Consciousness Studies*, *29*(1-2), 56-78.

[18] Seth, A. K. – Bayne, T. (2022). “Theories of consciousness.” *Nature Reviews Neuroscience* 23/7, 439-452.

[19] Tononi, G. (2012). “The integrated information theory of consciousness: an updated account.” *Archives italiennes de biologie*, *150*(2/3), 56-90.

[20] Tononi, G. (2015). “Integrated information theory.” *Scholarpedia*, *10*(1), 4164.

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

[22] Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). “The ConTraSt database for analysing and comparing empirical studies of consciousness theories.” *Nature Human Behaviour*, *6*(4), 593-604.