JNphi

Integrated Information Theory 4.0 is both Weakly Panpsychist and Strongly Dualist, but many Theories of Consciousness are also Prone to It

Sergey B. Yurchenko

Abstract

Since its first formulation the Integrated Information Theory (IIT) has recently been updated to the version 4.0. Unlike the previous versions where the problem of free will was completely neglected, IIT 4.0 claims to suggest a full neuroscientific account of this oldest problem in the philosophy of mind. The aim of this opinion paper is to show that IIT's account of free will is apparently dualist and reminiscent of the conventional free will in folk psychology, where mental constructs such as beliefs and desires are regarded as actual causes of human actions. On the other hand, these mental constructs can have high predictive power, compared to that provided by neuroscience. Thus, while rejecting ontological dualism, one can accept *methodological* dualism, compatible with eliminative physicalism, by virtue of its predictive power and descriptive parsimony.

Key Words: consciousness, dualism, panpsychism, free will, causation

DOI: 10.5281/zenodo.10877331

Introduction

Many scientists and philosophers agree that consciousness is not an independent phenomenon but emerges from brain activity in causal ways. But they are quick to point out that consciousness is not a physical entity. Instead, it is a mysterious process that supervenes on physical events. Although those events are part of the material world, which is causally closed, this opens a loophole for the complementarity of mind and matter, which, they think, differs from dualism.

The newest version of Integrated Information theory (IIT) makes it explicitly by dividing reality we live in into two forms of existence: *intrinsic* existence for conscious entities, and *extrinsic* existence for non-conscious entities (Albantakis et al., 2023). In its declarative part, IIT 4.0 aims to account for phenomenal properties of subjective experience in physical terms which can be observed and manipulated. Its starting point is the existence of experience, which is an immediate and irrefutable truth as it

Corresponding author: Sergey B. Yurchenko
Address: Brain and Consciousness Independent Research Center, P.O. 710132, Andijan, Uzbekistan
e-mail ✉ s.yucko@gmail.com

is encapsulated in Descartes' "Cogito ergo sum" (I think, therefore I am). This takes the form of the zeroth axiom:

(0) Existence Experience exists: there is something.

On this basis, IIT 4.0 IIT identifies the following five axioms:

(1) Intrinsicality Experience is intrinsic: it exists for itself.
(2) Information Experience is specific: it is this one.
(3) Integration Experience is unitary: it is *a whole*, irreducible to separate experiences.
(4) Exclusion Experience is definite: it is *this whole*.
(5) Composition Experience is structured.

By characterizing physical existence operationally, IIT takes then a strange leap from the axioms of *phenomenal* existence to the postulates of *physical* existence translated into the physical substrate of consciousness. The most fundamental and relevant to this opinion paper is the zeroth postulate (Albantakis et al., 2023): The substrate of consciousness can be characterized operationally by cause–effect power: its units must take and make a difference (p.5). The leap becomes especially remarkable in the formulation of Tononi et al. (2022) as the "principle of being": To exist physically means to have cause–effect power (p.3).

This transition from the zeroth axiom to the consequent postulate conflates two different notions, which are indeed doubtless separately but not the same one together: phenomenal existence in Cartesian sense, and physical existence in terms of 'normal' metaphysics. We agree that the existence of experience is given to us immediately, without reference to other evidence. We also agree that the principle of being is a kind of tautology: what does not exist physically, e.g., mental abstractions such as numbers, geometrical figures, or ghosts, cannot have causal power in the physical world. Consciousness exists *phenomenally* from its own intrinsic perspective, while the brain (or the embodied mind) exists *physically* from an observer's extrinsic perspective. How can it account for free will?

Panpsychism and Dualism in IIT 4.0

In neuroscience, free will is typically studied in experimental setting as a self-initiated action under the control of consciousness. In the philosophy of mind, free will refers to mental causation which is in apparent tension between mind and matter. To solve the tension, IIT 4.0 takes for the antecedent what must be just proven, namely the causal power of consciousness, which phenomenal existence has then been turned into its physical existence by the zeroth postulate.

The argument is this: IIT's argument for true free will hinges on the proper understanding of experience as true existence and on the intrinsic

powers view: what truly exists, in physical terms, are intrinsic entities, and only what truly exists can cause (Tononi et al., 2022, p.19).

Its underlying reason is bewildering: The ultimate reason is that as a conscious being, I truly exist and truly cause, whereas my neurons or my atoms neither truly exist nor truly cause (p.2).

What we have here is that IIT's proof of free will is either inconsistent as being logically circular, or it does not prove free will at all but only postulates its existence by means of involving dualist ontology. Note that Descartes himself did not equate his famous Cogito, formulated in terms of logic as a self-evident truth of introspection, with Cartesian dualism, just *postulated* metaphysically to justify Cogito, i.e., the unprovable existence of our subjective experience in the physical world full of inanimate things. Unlike us, these things exist physically but lack conscious experience, or there should be adopted panpsychism.

The proponents of panpsychism such as Leibniz and Spinoza endorsed the view that conscious properties are readily instantiated in matter. In contrast, Descartes had suggested dualism. These both have their own merit. Panpsychism is the most elegant answer to the hard problem of consciousness (Chalmers, 1989): how and why the mental is incorporated into the physical. Dualism proposes a simple solution to the problem of free will: how and why the mental can causally affect the physical.

But IIT 4.0 attempts to 'kill two birds with one stone'. Obviously, panpsychism and dualism would not stand alone over centuries, and these two problems should not be so difficult, if they were solvable in a unified way, proposed by Tononi and colleagues. In fact, whenever a theory speaks about a volitional role of consciousness, the theory implicitly admits dualism as if consciousness might exist independently of neural activity to control the brain in making its work. And IIT 4.0 makes this claim unambiguously by diving the ontological basis of the world into two parts, called "the great divide of being" (Tononi et al., 2022, p.8): the one – for the intrinsic existence of consciousness, associated with the maxima of integrated information, a system is capable of generating, and the other – for the extrinsic existence of its physical substrates that can be observed and manipulated.

The metaphysical confusion, produced by IIT 4.0, goes on when the theory insists on its commitments to realism, physicalism, and reductionism to account for the (extrinsic) existence of atoms and neurons that are an indispensable material basis of subjective (intrinsic) experience. It is not surprisingly, therefore, that some authors (Cea et al., 2023) call IIT an idealist theory, which is in tension with the realist assumption that the world of things exists independently of our being. On the other hand, Mørch (2019) argues that the axiom of intrinsicality is incompatible with reductionism, which deprives consciousness of primacy over matter.

After all, a letter, signed by 124 neuroscientists, psychologists, and philosophers, argues that IIT's commitments to panpsychism makes it pseudoscience (Fleming *et al.,* 2023). Although the accusation is made from the perspective of IIT's panpsychist impact on clinical practice and

ethical issues, concerning coma patients, fetuses, and brain organoids, it is odd by two reasons. First, there is now a great number of different theories of consciousness (ToCs) (Seth and Bayne, 2022; Evers et al., 2024), and many of them are implicitly prone to some degree of panpsychism (Lamme, 2018) as far as they strive to arrive at a universal definition of consciousness that could be applicable to different physical systems, not only to brain-centric organisms (Kanai and Fujisawa, 2023). Second, in contrast to other ToCs, since its early version in Manifesto (Tononi, 2008), IIT has explicitly emphasized its panpsychist flavor as a merit that can help to solve the hard problem of consciousness, i.e., to explain scientifically not only the objective – behavioral, functional, and neural correlates of consciousness, but also its subjective – intrinsic properties (Ellia et al., 2021).

In general, the panpsychist flavor of IIT appears spontaneously from its claim that the maxima of integrated information, called Φ (Phi), are uniquely identical to conscious experience a subject has at that time (Tononi and Koch, 2015). Thus, any system – natural or artificial, brain-centric or not – that is capable of generating $\Phi_{max} > 0$, will be conscious. On the other hand, a system that does not integrate information more than its parts cannot be conscious. Thus, IIT is weakly panpsychist insofar as it does not imply that consciousness is ubiquitous in the universe.

In contrast, IIT's strong form of dualism is the main premise for the irreducible cause-effect power of Φ_{max} to account for free will. The commitments to realism and reductionism do not allow IIT to admit that consciousness, associated with Φ_{max}, can act by itself in physical world. Instead, IIT makes consciousness causal powerful in how it can affect brain activity via mental downward causation.

What follows is that IIT's account of free will is implicitly based on the two assumptions (Yurchenko, 2023a):

1. Downward causation across spatial scales is possible;
2. Information has causal power over and above that provided by matter.

It is well known that the first assumption entails overdetermination (Kim, 2016) or, more exactly, the double causation fallacy, when two causes, represented by a microscopic (physical) variable and by a macroscopic (coarse-grained or supervenient) variable, are responsible for the same effect. The second assumption implies mental causation and requires the mind-body dualism, defined as the "great divide of being" (Fig.1).

Of course, the free will problem is not specific for IIT exclusively. This is a general issue for all ToCs. No theory claiming to explain subjective experience – be it based on the integrated information, global workspace, predictive (Bayesian) processing, macroscopic quantum entanglement, self-organized criticality, or on something else – can sidestep this problem. And many ToCs, while being very different in how they define consciousness, converge to the idea that consciousness should have a causal function in brain dynamics. Its active role can be

linked there to broadcasting (Dehaene and Naccache, 2001), to active inference (Friston et al., 2013), or to metacognition (Rosenthal, 2008). At any case, all these ToCs can be characterized as covertly dualist by involving the assumption 1 and 2 in their foundations.

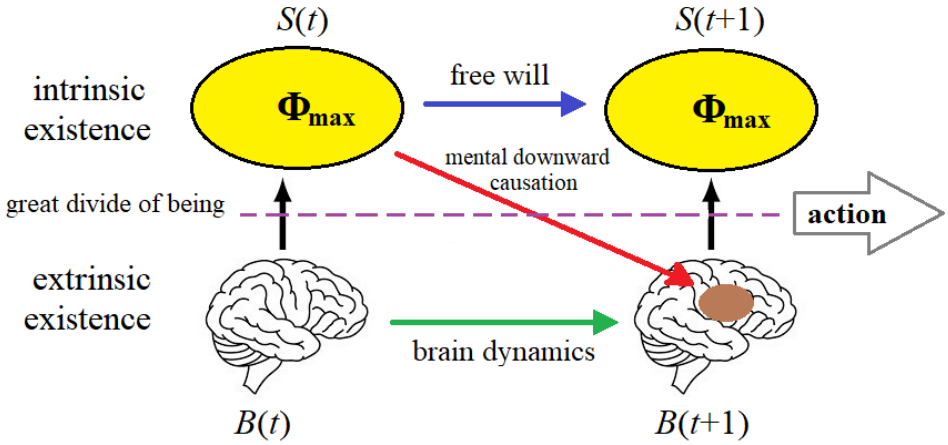

Figure 1. Here, consciousness is the stream of transient states, each presented by a supervenient variable $S(t)$ and identified with Φ_{max}. Formally, the stream is a transitive and irreflexive (Markovian) chains of conscious states, emerging (black arrow) from the corresponding brain states $B(t)$, which all evolve causally on their own (green arrow) over time. Free will is defined as the ability of consciousness to choose its next state $S(t+1)$ from the previous one (blue arrow). Thus, an action, associated with this state, might be implemented by affecting certain brain regions, e.g., motor modules (brown circle), via mental downward causation (red arrow). There is double causation between brain dynamics and mental causation.

Consciousness in Neuroscience and Folk Psychology

Consciousness is a "mongrel" concept that connotes a number of different notions (Block, 1995). In folk psychology, consciousness is commonly associated with the self, referred in everyday language to as human 'I'. In this framework, consciousness is extended over time and has beliefs, desires, hopes, and intentions, which all are commonly regarded as actual causes of human actions. In contrast, in neuroscience, consciousness is nothing more than the stream of conscious states (Yurchenko, 2023b) that vanishes in sleep, in epileptic seizure, and in coma. The obvious inference drawn from it is that consciousness is an emergent (serial) phenomenon, having no causal power over brain activity. Before choosing which action to perform the brain should spontaneously initiate a decision-making process. Once the decision has been finalized and simultaneously exposed to the conscious level, the self-initiated action can be performed, as it is abundantly detected in Libet-type experiments (Libet, 1985; Soon et al., 2013; Khalighinejad et al., 2018).

However, Tononi et al. (2022) make their general statement concerning free will in everyday language of folk psychology:

> I can have true free will: I can have true alternatives, true freedom to choose among them, true will to cause what I have decided, and eventually true responsibility (p.2).

But who am I if not what my brain is doing just now by generating the stream of conscious states? The brain processes information not I, experiencing its outcomes which give me the sense of agency over time. The arguments against this conclusion, drawn from Libet-type experiments, are usually based on the idea that consciousness, i.e., the self, can still be free in making deliberate decisions (Maoz et al., 2019) and in pursuing its distal rather than proximal intentions (Nichelli and Grafman, 2024). However, if consciousness cannot choose its next state $S(t+1)$, how might consciousness be free to make deliberate decisions in choosing its future state $S(t+n)$ that was also preceded by the previous state $S(t+n-1)$. The mechanism of binding two nearest states in the stream, generated by the brain time after time, must be universal (Yurchenko, 2022). The brain cannot concoct serial conscious states ad hoc. It is also crucial for our understanding of free will that consciousness cannot move back and forth over the stream. This is what it is right now. When viewed through the lens of this framework, the dualist foundations of folk psychology (Wisniewski et al., 2019), covertly involving the assumptions 1 and 2 to account for the causal power of mental constructs, become apparent.

The explanation why covert dualism of folk psychology is so attractive for laypeople and for some scientists lies in its simplicity. In mathematical models of cognitive and behavioral neuroscience, experimental psychology and social sciences, consciousness can be generally characterized as a supervenient variable. *Supervenience* here means the relationship between two classes of variables (properties), where one class is more fundamental than the other so that the variables of the upper class emerge from, or are determined by, the variables of the lower class. A typical example in the philosophy of mind is mental properties that are said to supervene on physical properties.

Neuroscience deals with a great number of different variables at a micro-, a meso-, and a macro-scale for describing brain activity. Yet, there happen involved supervenient variables, associated with perception, cognition, and consciousness properly. The supervenient variables are weakly emergent in the sense that they are macroscopic, as emerging over a system of interest, and "somehow autonomous from underlying processes" but do not have irreducible causal power over the system (Bedau, 1997). In other words, these variables can provide high predictive power about their own dynamics, but they cannot have causal effects on the lower microscopic variables. Otherwise, double causation would be involved (Fig.1).

Ultimately, all these details make neuroscientific models very complex, computationally difficult, yet often opaque for interpretation. In contrast, the concepts of folk psychology such as beliefs, desires, and intentions, though not being identifiable by particular conscious states, are intuitively understandable and deceptively simple at first sight. Nonetheless, these mental abstractions are eliminated from

neuroscientific descriptions as not being amendable to further analysis. Indeed, if even some belief (in what precisely?) was somehow specified, what neural basis should underlie it?

Newsome (2014) asks:

> The critical question is whether our beliefs, values, and aspirations … are real entities with real causal efficacy in the world or whether they are illusory constructs that we make up to describe our experience of a world whose causal determinants lie at a much more fundamental level (p.94).

He points out three advantages of folk psychology: predictability, manipulability, and parsimony, and argues that mental constructs are organizational entities, instantiated in high-level neural systems within the brain, which resist explanation through eliminative reduction.

> Considering humans to have real mental states with causal efficacy has overwhelming advantages for predicting the future… Criteria of manipulability, in addition to prediction, argue for the validity of minds as real, causal entities (Newsome, 2014, p.91).

As stated, brain dynamics embraces a range of scale-dependent variables, including supervenient variables, and each of them can have predictive power, yet be manipulated (perturbed). The same is true for mental constructs. Knowing a human's habits can help in predicting her reaction on a piece of information (e.g., whether she will take one action over the other in experimental settings) to a degree that might not be reached from modeling at the neural level.

Thus, predictions, derived from the mental abstractions of folk psychology, without knowing anything about the underlying brain activity, can sometimes surpass the power of neuroscientific predictions. This makes folk psychology a good tool for *cause-like* explanations. Nonetheless, such explanations could scarcely help us to uncover the genuinely physical causes of neural abnormalities in mental illnesses, e.g., in schizophrenia (Takayanagi et al., 2021), accompanied with the false *belief*s (delusions) folk psychology relies upon. How might its predictive power, manipulability, and parsimony advance our understanding and medical treatment of schizophrenia?

Concluding Remarks

In general, covert dualism of folk psychology can have methodological advantages over causal descriptions in neuroscience. This kind of dualism has already been widely accepted in many branches of science, in biology, in social sciences, in economics, where dominate population dynamics (e.g., prey-predator models) or game-theoretic models, all describing individuals as free agents acting to attain their preplanned, mentally-driven goals. Methodological dualism can take the legitimized form of *reification* or "the fallacy of misplaced concreteness" (Whitehead, 1978), when abstract constructs such as mental intentions or, more broadly, biological purposefulness are regarded as if they were real physical events, having both predictive and causal power over a system of interest.

In neuroscience, many ToCs are also prone to this methodological dualism at risk of falling into its ontological form. The IIT's account of free will is, probably, the best example of taking it ontologically.

References

Albantakis L, Barbosa L, Findlay G, Grasso M, Haun AM, et al. Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms. 2023. arXiv:2212.14787

Bedau MA. Weak emergence. Nous 1997; 31(11):375–399.

Block N. On a confusion about a function of consciousness. Behav Brain Sci 1995; 18:227-287.

Cea I, Negro N, Signorelli CM. The Fundamental Tension in Integrated Information Theory 4.0's Realist Idealism. Entropy 2023; 25:1453.

Dehaene S, Naccache L. Towards a cognitive permanence of consciousness: basic evidence and a workspace framework. Cognition 2001; 79:1–37.

Ellia F, Hendren J, Grasso M, et al. Consciousness and the fallacy of misplaced objectivity. Neurosci Conscious 2021; 7(2):1–12. niab032

Evers K, Farisco M, Pennartz CMA. Assessing the commensurability of theories of consciousness: On the usefulness of common denominators in differentiating, integrating and testing hypotheses. Conscious Cogn 2024; 119:103668.

Fleming SM, Frith C, Goodale M, et al. The Integrated Information Theory of Consciousness as Pseudoscience. 2023; https://doi.org/10.31234/osf.io/zsr78

Friston KJ, Schwartenbeck P, FitzGerald T, et al. The anatomy of choice: active inference and agency. Front Hum Neurosci 2013; 7:598.

Kanai R, Fujisawa I. Towards a Universal Theory of Consciousness. 2023; https://doi.org/10.31234/osf.io/r5t2n

Khalighinejad N, Schurger A, Desantis A, et al. Precursor processes of human self-initiated action. Neuroimage 2018; 165:35–4.

Kim J. Emergence: Core ideas and issues. Synthese 2016; 151 3:547–59.

Libet B. Unconscious cerebral initiative and the role of conscious will in voluntary action. Behav Brain Sci 1985; 8:529–566.

Maoz U, Yaffe G, Koch C, Mudrik L. Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. eLife 2019; 8:e39787.

Mørch HH. Is consciousness intrinsic: A problem for the integrated information theory. J Conscious Stud 2019; 26:133–162.

Newsome WT. Neuroscience, Explanation, and the Problem of Free Will. in W. Sinnott-Armstrong (ed.), Moral Psychology: Free Will and Moral Responsibility. Vol. 4. Cambridge: MIT Press 2014; pp. 81–96.

Nichelli PF, Grafman J. The place of Free Will: the freedom of the prisoner. Neurol Sci 2024; 45(3):861-871.

Rosenthal DM. Consciousness and its function. Neuropsychologia 2008; 46:829–840.

Seth AK, Bayne T. Theories of consciousness. Nat Rev Neurosci. 2022; 23(7):439-452.

Soon CS, He AH, Bode S, Haynes JD. Predicting free choices for abstract intentions. Proc Natl Acad Sci U S A 2013; 110:6217–6222.

Takayanagi Y, Ishizuka K, Laursen TM, et al. From population to neuron: exploring common mediators for metabolic problems and mental illnesses. Mol Psychiatry 2021; 26(8):3931-3942.

Tononi G. Consciousness as integrated information: a provisional manifesto. Biol Bull 2008; 215:216–242.

Tononi G, Koch C. Consciousness: here, there and everywhere? Phil Trans R Soc London, B. 2015; 370:20140167.

Tononi G, Albantakis L, Boly M, et al. Only what exists can cause: An intrinsic view of free will. 2022; https://arxiv.org/abs/2206.02069

Whitehead AN. Process and Reality. 1978; New York: The Free Press.

Wisniewski D, Deutschländer R, Haynes JD. Free will beliefs are better predicted by dualism than determinism beliefs across different cultures. PLoS One 2019; 14(9):e0221617.

Yurchenko SB. From the origins to the stream of consciousness and its neural correlates. Front Integr Neurosci 2022; 16:928978.

Yurchenko SB. Is information the other face of causation in biological systems? BioSystems 2023a; 229. https://doi.org/10.1016/j.biosystems.2023.104925

Yurchenko SB. A systematic approach to brain dynamics: cognitive evolution theory of consciousness. Cognitive Neurodynamics 2023b; 17, 3:575-603.