Dear Scott,

I recently completed my Ph.D. at Caltech under Christof Koch (CK) immersing myself in the nitty-gritty of IIT for the past 6.5 years. After reading the discussion among you, David Chalmers, and Giulio Tononi (GT). I feel there's several places I can supplement the discussion. I apologize for taking so long to respond, but I figure better late than never.

First, thank you for your IIT posts; they have extended IIT-awareness and been productive in IIT theory-crafting. Ultimately, I agree with many of your points, however, there's some bits I believe will improve IIT's defense of your welcomed criticisms.

Here I'll respond to your *Why I'm not an Integrated Information Theorist* quoting the relevant sections and providing feedback. In the next letter I'll respond to *Giulio Tononi and me: A Phi-nal Exchange.*

> *(2) to hypothesize that a physical system is "conscious" if and only if it has a large value of ϕ—and indeed, that a system is more conscious the larger its ϕ value.*

The "if and only if" is debatable among the IIT proponents. I personally see phi as taking unambiguous properties of phemonological experience, recasting them as equations, and then asserting these equations are necessary conditions for consciousness (C). Naturally, this process can be iterated to establish increasingly stringent preconditions for C.

While in his lab CK explicitly stated (verbally) that he viewed phi as a *necessary* condition for C and was agnostic about whether it is a sufficient condition. However, in the most recent communication he stated that he viewed ϕ as also sufficient for C. I wasn't able to determine the reason for the change of opinion.

GT almost always implies, and sometimes explicitly says that positive phi is necessary and sufficient for C. However, it's murky how to reconcile this claim with the widely different versions of ϕ. For example, phi-2004 [1] has no notion of time, phi-2008 [2] looks "backwards" in time, and phi-2014 looks both backwards and forwards! Another example is the addition of the exclusion axiom which was added in phi-2014 [3]. Which one is sufficient for C? All of them? Phi-2008? Only phi-2014? If only phi-2014, why that one? If the phi-2008 was sufficient for C, then why the new axiom and postulates?

When GT had previously implied phi-2008 was sufficient for C, I presumed he was simply being socially provocative—it's how you get your theory talked about. If GT sincerely believes positive phi is sufficient for C, the only way I can imagine this argument working is to iterate the above process such that each successive phi-measure takes into account additional phenomological properties of C. Taking the limit of this process yields a phi-measure which incorporates every apparent phenomological feature of conscious. And then you could assert that this is "all that needs explaining" for a theory of C. As-is, there has been no argument for why the existing axioms of differentiation, integration, and exclusion fully exhaust the phenomological properties requiring explanation.

To move IIT from talked about to accepted among hard scientists, it may be necessary for GT to wash his hands of sufficiency claims.

> *Since I don't know a standard name for the problem, I hereby call it the Pretty-Hard Problem of Consciousness.*

Like David Chalmers, I too enjoy your description of the PHP, but I should clarify IIT's stance within Chalmer's PHP classes. Chalmers defines:

- PHP3: "Construct a theory that tells us which systems are conscious".
- PHP4: "Construct a theory that tells us which systems have which states of consciousness".

In phi-2008 they address PHP3.5, which I define as "Construct a theory that tells us the magnitude of a system's consciousness". It's often said that Shannon information theory tells us "the volume of a substance" (information) without ever saying "what the substance actually is". As such PHP3.5 is a fitting task for Shannon theory.

IIT flirts with PHP4 in [4] and even more so in phi-2014. And indeed specialists across information theory have for decades [5, 6, 7] attempted to go beyond quantifying the "volume" of information to understanding the structure of information itself. This remains very much an open problem[8, 9], and one researchers since von Neumann have spent careers on.[1]

Given this situtation, I doubt even IIT proponents put much faith in phi-2014's improvised, makeshift Earth-Mover's Distance solution to the "structure of information" problem—I certainly don't. However, the important thing is that IIT **does not** stop at PHP3.5—it very much aims to (eventually) tackle PHP4; the necessary mathematics for getting there just haven't been discovered yet.

> *In my view, IIT fails to solve the Pretty-Hard Problem because it unavoidably predicts vast amounts of consciousness in physical systems that no sane person would regard as particularly "conscious" at all...*

There's two responses to this. The easiest response is to say that ϕ is merely necessary for C—problem solved. GT's response would be to challenge your intuition for things being unconscious. Here's a historical analogy; imagine when the Kelvin temperature scale was introduced. Here Kelvin was saying that *just about everything* has heat in it. In fact, even the coldest thing you've touched actually has substantial heat in it! Think of IIT as attempting to put a Kelvin-scale on our notions of C. I find this "Kelvin scale for C" analogy makes the panpsychism much more palatable.

[1] See also the "conceptual structure" section in the next letter.

> *Strikingly, despite the large literature about ϕ, I had a hard time finding a clear mathematical definition of it—one that not only listed formulas but fully defined the structures that the formulas were talking about.*

You've placed your finger on a major problem in the IIT literature. IIT needs more mathematically inclined people at its helm. GT in particular desperately needs to start co-authoring with a mathematician or theoretical hard scientist who is not on his payroll. (Maybe you?)

> *Similarly, we define $EI(B \to A) := H(zA|Brandom, yA = xA)$.*

This isn't quite the right intuition nor the right expression. To start with this expression maximizes for a system that is completely uncorrelated with its past, i.e., $EI(B \to A) = H(zA) = H(zA|Brandom, yA = xA)$. Intuitively, you want to condition on a state of the *output* of the function and look at how much of the uncertainty about the input has been reduced, i.e., $EI(B \to a) = I(B_t : a_{t+1}) = H(B_t) - H(B_t|a_{t+1})$.

> *For this reason, Tononi proposes a fix where we normalize each $\phi(A, B)$ by dividing it by* $\min(|A|, |B|)$.

The normalization procedure in phi-2004 and phi-2008 are post-hoc hacks to make the resulting partition look more like modules. I argue they should both be discarded. If ϕ measures the (minimum) "irreducibility to any partition", then the bipartition $A = \{1, \ldots, n-1\}$ and $B = \{n\}$ is perfectly allowed. If getting highly asymmetric minimum partitions is unpleasant, then we are instead wishing to quantify irreducibility to *functional modules*. If it's irreducibility to modules we want, there are numerous more principled procedures [10] for finding modules than ad-hoc normalizations. (If we want to propose a module-finding algorithm using irreducibility-to-a-partition, that's a fine idea but is logically separate and should be evaluated separately.) If you have multiple modular partitions, you choose the one with the lowest unnormalized irreducibility/phi.

> *To be sure, empirical work in integrated information theory has been hampered by three difficulties....*

These are very good points. I do not know how to solve these practical matters. I have limited and contented myself to simply getting the theory into better shape.

> *As humans, we seem to have the intuition that global integration of information is such a powerful property that no "simple" or "mundane" computational process could possibly achieve it.*

You might like this transform, the All-or-nothing transform [11]. I would look here for nice examples yielding very high ϕ.

Sincerely,
Virgil Griffith
`virgil@caltech.edu`

References

[1] Tononi, G. (2004) An information integration theory of consciousness. *BMC Neurosci* **5**, 42.

[2] Balduzzi, D & Tononi, G. (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology* **4**, e1000091.

[3] Oizumi, M, Albantakis, L, & Tononi, G. (2014) From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput Biol* **10**, e1003588.

[4] Balduzzi, D & Tononi, G. (2009) Qualia: The geometry of integrated information. *PLoS Computational Biology* **5**.

[5] Burgin, M. (2009) *Theory of Information: Fundamentality, Diversity and Unification.* (World Scientific Publishing Company).

[6] Floridi, L. (2014) Semantic conceptions of information (http://plato.stanford.edu/archives/spr2014/entries/information-semantic/).

[7] Klir, G. J. (2004) Generalized information theory: aims, results, and open problems. *Reliability Engineering and System Safety* **85**, 21–38. Alternative Representations of Epistemic Uncertainty.

[8] Spivak, D. (2014) What is the underlying mathematical structure of information itself? (http://math.mit.edu/ dspivak/informatics/).

[9] Williams, P. L & Beer, R. D. (2010) Nonnegative decomposition of multivariate information. *CoRR* **abs/1004.2515**.

[10] Ziv, E, Middendorf, M, & Wiggins, C. H. (2005) Information-theoretic approach to network modularity. **71**, 046117.

[11] (2014) (https://en.wikipedia.org/wiki/All-or-nothing_transform).