Scott,

This is my feedback to your post, *Giulio Tononi and me: A Phi-nal Exchange.* In Giulio Tononi's (GT) response to you, he writes:

> *Scott considers at some length the Vandermonde matrix. As he himself realizes, this is an abstract mathematical entity, and IIT deals explicitly with actual physical systems, not mathematical idealizations, so Vandermonde is not directly relevant.*

If positive phi is already sufficient for consciousness (C), I don't see how one can justifiably append a new precondition, in this case, "physicalness". If physicalness is a requirement for C, and positive phi is sufficient for C, then ϕ should be in physical units, e.g., joule-bits per cubic-meter.On the other hand, if we said that phi was merely a necessary condition for C, this criticism disappears because you can then simply assert physicalness as an additional precondition which the phi measure doesn't (yet?) include.

Your Phi-nal response to GT

> *The conceptual structure is unified — it cannot be decomposed into independent components (integration).*

Minor point—the more precise term is that the structure cannot be *partitioned* into independent components. The term "decomposed" allows overlapping components; GT explicitly forbids these because they violate the components' "physicalness". You can see examples demonstrating these various notions of irreducibility (e.g., overlapping vs non-overlapping parts) in [1].

> *the definitions never seemed to "bottom out" in mathematical notions that I understood, like functions mapping finite sets to other finite sets. What, for example, is a "mechanism"? What's a "system of mechanisms"? What's "causal power"? What's a "conceptual structure," and what does it mean for it to be "unified"?*

Agreed—this is a big problem with the IIT literature. A few of these I can define for you.

- A "mechanism" is a discrete conditional probability distribution representing an update rule. I.e., if you had atomic nodes $\{1, \ldots, n\}$ and X_1 is a random variable denoting the states of node 1 before the update rule, and Y_1 is a random variable denoting the states of node 1 after the update rule. Then the "mechanism" of node 1 is the condition probability distribution $p(Y_1|X_1)$.

- A "system of mechanisms" is simply the joint conditional distribution $p(Y_1 \ldots Y_n | X_1 \ldots X_n)$. Note this must be the full *joint* conditional distribution, not the set of mechanisms $\{p(Y_1|X_1), \ldots, p(Y_n|X_n)\}$.
- The term "causal power" means different things to different people. But (as far as I can tell) when GT says "causal power", he means the *probability of necessity* (PN) from the Pearl framework (see http://bayes.cs.ucla.edu/BOOK-2K/ch9-1.pdf for an intro). To get a PN-distribution into a nonnegative scalar representing "causal power", you need to plug the PN-distibution into one of measures given in [2, 3]. Although quantifying causal power is much closer to a solved problem than the internal structure of information (see next bullet), quantifying causal power remains an area of ongoing research and the measures will certainly be further refined.
- As for "conceptual structure", GT intends this to be the topology of information passing through the update rule. The established way for representing this "constellation" topology GT is looking for is Grothendieck topology. The best person I know working rigorously with these kinds of structures, i.e., understanding the "internal structure of information", is David Spivak in the E17 building. I don't know if Prof. Spivak knows about IIT, but he is the most able person I know of to represent GT's "conceptual structure" using rigorous, non-idiosyncratic methods.
- The term "unified" is the trickiest to know what GT intends. As far as I can tell he uses this term as a synonym for "integration with a hint of the exclusion axiom", and I spent my Ph. D. focusing on rigorizing this one concept [4]. I went through several notions [5, 6, 7, 8], but I eventually settled on the informational synergy framework from [9]. Even if [9]'s notion of synergistic mutual information isn't quite what GT means by "unified", it is definitely in the conceptual ballpark and moreover is something interesting and indepedently important in computational neuroscience and genetics[10, 11]. The perfect expression for calculating synergistic mutual information per [9] is an open problem, but the bounds ever tighten [12, 13] and will probably be solved within a few years.

> *Instead, there was general discussion of the postulates, and then ϕ just sort of appeared at some point.*

Agreed 110%. This is a pothole in the theory. Collaborating with hard scientists not on his payroll could greatly smooth this over.

> *Furthermore, given the many idiosyncrasies of ϕ—the minimization over all bipartite (why just bipartite? why not tripartite?) decompositions of the system, the need for normalization (or something else in version 3.0) to deal with highly-unbalanced partitions*

Also agreed 110%. In fact these same exact issues bothered me so much that I aimed to answer them betore worrying about things like causal power or practicalities. The result of my work is the ψ measure [14]. In my version, I prove a bipartition will always have minimum synergy and explicitly disavow any normalization as a hack. I view ψ as a more principled version of phi-2008 (and thus only aspires to answer PHP3.5) and suspect you'll find it much more agreeable to your tastes than phi-2004 [15], phi-2008 [16], or phi-2014 [17].

I personally would be happy sticking to getting PHP3.5 down solid before reaching for PHP4. But for those interested in PHP4, the first thing would be to replace the jury-rigged Earth-Mover's-Distance in phi-2014 with something from Grothendieck topology.

> *Well, if the postulates uniquely determined the form of ϕ, then what's with all these upgrades? Or has ϕ's definition been changing from year to year because the postulates themselves have been changing?*

The short answer is, "both have been changing". The easy examples are adding the exclusion axiom in phi-2014 [18, 17] and the progression from phi-2004 to phi-2008 to phi-2014 in how each one treats time radically differently.

To make matters worse, you'd at least hope the phi axioms/postulates would define the **units** for phi. But unfortunately it does not. Phi-2004 and phi-2008 are in units of "bits", phi-2012 [18] is in units *bits-squared*, and phi-2014 has no units. Rome wasn't built in a day, and in GT's response to you he concedes it remains unclear how to treat time. Moreover, even as a work-in-progress IIT remains quite useful! However, phi's claims of sufficiency for C (since 2008) are unbecoming when future versions can't even agree on the units.

> *Clearly, a theory of consciousness must be able to provide an adequate account for such seemingly disparate but largely uncontroversial facts. Such empirical facts, and not intuitions, should be its primary test...*

Solid. This is exactly what an IIT proponent would say.

> *you can't count it as a "success" for IIT if it predicts that the cerebellum in unconscious, while at the same time denying that it's a "failure" for IIT if it predicts that a square mesh of XOR gates is conscious.*

To ensure I understand you correctly, you are saying that the implausibility of the cerebellum lacking its own C separate from ours is comparable to the implausibility of an XOR gate-mesh having high-magnitude C. Ergo, the two roughly "cancel out" for the validity of IIT.

I see two ways an IIT-proponent can respond:

1. From phenomenology, use the integration axiom as a necessary condition for C. Then, since the cerebellum's neural structure has low integration (for any reasonable definition of integration), then the cerebellum has at most modest C.
2. From phenomenology, use the exclusion axiom as a property of C. Then, since the cerebellum is a component our larger brain, then cerebellum is precluded from having its own C (under the weakest form of the exclusion axiom, the cerebellum at least doesn't has lower C than us).

Using either of these sketches of arguments, the IIT proponent can make the cerebellum's low-C more plausible than the XOR network's high-C. Therefore, an IIT-proponent would say that predicting the cerebellum's low-C "weighs more" than the counter-intuitive prediction that the XOR mesh has high-C.

> *Now, it would be child's-play to criticize the above line of argument for conflating our consciousness of the screen with the alleged consciousness of the screen itself. To wit: Just because it feels like something to see a wall, doesn't mean it feels like something to be a wall. You can smell a rose, and the rose can smell good, but that doesn't mean the rose can smell you.*

This is a miscommunication. Even when both are inebriated, I've never heard GT nor CK separately or collectively imply anything like this. Moreover, they're each far too clueful to fall for something so trivial. For whatever it's worth, GT is usually clearest in person and CK is usually clearest in writing.

I hope this improved understanding of IIT improves its standing,
Virgil Griffith
`virgil@caltech.edu`

References

[1] Griffith, V & Harel, J. (2013) Irreducibility is minimum synergy among parts. *CoRR* **abs/1311.7442**.

[2] Korb, K. B, Hope, L. R, & Nyberg, E. P. (2009) in *Information Theory and Statistical Learning*. (Springer), pp. 231–265.

[3] Dominik Janzing, David Balduzzi, M. G.-W & Schoelkopf, B. (2012) Quantifying causal influences. *arXiv:1203.6502*.

[4] Griffith, V. (2014) Quantifying synergistic information (http://thesis.library.caltech.edu/8041/).

[5] Griffith, V & Koch, C. (2014) *Quantifying synergistic mutual information* ed. Prokopenko, M. (Springer).

[6] Gell-Mann, M & Lloyd, S. (1996) Information measures, effective complexity, and total information. *Complexity* **2**, 44–52.

[7] Zurek, W. H, ed. (1990) *Complexity, entropy, and the physics of information*, SFI Studies in the Sciences of Complexity. (Addison-Wesley) Vol. 8.

[8] Prokopenko, M, Boschetti, F, & Ryan, A. J. (2009) An information-theoretic primer on complexity, self-organization, and emergence. *Complexity* **15**, 11–28.

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

[10] Narayanan, N. S, Kimchi, E. Y, & Laubach, M. (2005) Redundancy and synergy of neuronal ensembles in motor cortex. *The Journal of Neuroscience* **25**, 4207–4216.

[11] Varadan, V, Miller, D. M, & Anastassiou, D. (2006) Computational inference of the molecular logic for synaptic connectivity in c. elegans. *Bioinformatics* **22**, e497–e506.

[12] Griffith, V, Chong, E. K. P, James, R. G, Ellison, C. J, & Crutchfield, J. P. (2013) Intersection information based on common randomness. *ArXiv e-prints.*

[13] Rauh, J, Bertschinger, N, Olbrich, E, & Jost, J. (2014) Reconsidering unique information: Towards a multivariate information decomposition. *CoRR* **abs/1404.3146**.

[14] Griffith, V. (2014) A principled infotheoretic phi-like measure. *CoRR* **abs/1401.0978**.

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

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

[17] 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.

[18] Tononi, G. (2012) The integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie* **150**, 290–326.