What is Integrated Information Theory a Theory *Of?*[*]

Adam Pautz
December 2015
Brown University

Integrated information theory is supposed to be a theory of "the amount of consciousness in a system". When a system's Φ-value (a measure of "integrated information" in the system) is above 0, the system is conscious; and the value of Φ determines precisely the "amount" of consciousness it has (Tononi and Koch).

Many have noted that IIT has weird predictions - Scott Aaaronson especially has pressed this point. For instance, it implies that, if an extremely simple 2D grid has a Phi value that is (say) 10 times greater than your brain, then this 2D grid has 10 times the "amount" of consciousness that you have.

I agree that IIT has weird predictions. But if a theory of consciousness fits the data and is more elegant than the alternatives, maybe we should accept the theory even if it has some weird predictions. After all, some of our best physical theories have weird predictions too.

My central concern about IIT is different. I don't have a clear grasp on what the theory is a theory of. If you look at how IIT is formulated, it is not just a theory of when consciousness is present or absent. It is more specific; it is a theory of the *amount* of consciousness in an arbitrary system. The theory is that the *amount* of consciousness in a system is its level of Phi. So, for instance, it implies that, if a 2D grid has a Phi value that is (say) 10 times greater than your brain, then this 2D grid has 10 times the "amount" of consciousness that you have (even when you are fully awake and have had your morning coffee). Indeed, it implies that the amount of consciousness in such a system is *unbounded* - since its Phi level is unbounded.

[*] These remarks appeared in the discussion section of an article by John Horgan at *Scientific American* ("Consciousness and "Crazyism": Responses to Critique of Integrated Information Theory", December 7, 2015.) More recently, other have raised basically the same problem, namely that talk of "levels of consciousness" or "amount of consciousness" has been given no clear meaning (however they do not use this, as I have done, as an objection to IIT). See Bayne, T., Hohwy, J., & Owen, A. M. (2016). "Are There Levels of Consciousness?. *Trends in Cognitive Sciences*. 20(6), 405–413.

My worry about this is not Aaaronson's – namely, that such predictions are counterintuitive. Rather, my point is that it is not even clear what these predictions mean. What could it even mean to say that a 2D grid might have, say, "10 times the amount" of consciousness that you have when you are fully awake? My worry is not that this prediction is false; rather, *I don't know what this form of words even means*, so I cannot evaluate for truth or falsity. In general, I don't yet know what proponents of IIT mean by talk of the "amount" of consciousness – a supposedly unbounded dimension of our experiences (and indeed one that has a ratio scale, on IIT, since Phi has a ratio scale). Since "IIT" is supposed to be a theory of the "amount" of consciousness, and since proponents don't give this term a clear meaning, they haven't really specified a theory yet that can be evaluated.

By the "amount" of consciousness in a system, do they mean the number of experiences it has? Or the *intensity* of its experiences – so that if you turn up the volume on the radio, the "amount" of consciousness you are enjoying goes up? I am sure that they mean neither of these things. (They would not say, for instance, that the 2D grid has 10 times *the number* experiences than you, or that it has auditory experiences that are 10 times "louder" than yours, or anything of the sort.) Or do they perhaps mean the "amount" of information represented by an experience? But all experiences – even the experience of a blank wall – rule out infinitely many possibilities. Finally, by the "amount" of consciousness in a system, do they mean something about how much information that is being represented by conscious experience is being cognitively accessed (so that when you just wake up and are inattentive, you count as having a low amount of consciousness)? But this can't be what they mean either. For one thing, their view implies that there can be a large amount of consciousness even in a system, such as the 2D grid, *where there is no cognitive access at all*.

It might be replied that we find it hard to know what "amount of consciousness" means because we are stuck with our own "amount of consciousness". But this doesn't really address the problem. I have distinguished between different candidate meanings for "amount of consciousness". The proponent of IIT needs to say which one she means; until then her theory just hasn't been specified – she hasn't said what the theory is a theory of. The problem is that none of the candidates is plausible.

Another, separate problem with IIT, it seems to me, is that proponents discussed consciousness at a very abstract level. If we try to get down to the details, it is hard how the theory might explain even very rudimentary facts about experiences and their phenomenal structure. To take just one example: the intensity (perceived loudness) of your experience of one tone

might be twice greater than your experience of another tone. This is a phenomenological fact about your conscious experiences. But, given just the resources of IIT (Φ-value, nodes, cause-effect powers, etc.), it is very hard to see what this "doubling" might consist in or be grounded in (Φ-value doesn't double, "cause-effect powers" don't double!). If the theory lacks the resources to explain even this rudimentary phenomenological fact – it cannot even offer a candidate explanation – then it really hasn't made it out of the gate, or so it seems to me.