Prof. Searle, in your book review of Christof Koch's *Consciousness: Confessions of a Romantic Reductionist* in the New York Review of Books, you broach important criticisms of the Integrated Information Theory of consciousness (IIT). As someone who has worked in the nitty gritty of IIT for almost six years, I wanted to respond to your criticisms.

Common accord. Before getting to your two central objections to IIT, thank you for clarifying my own thinking. First, that consciousness is ontologically-subjective yet its existence is observer-independent. Second, that a proper science of consciousness must be epistemically-objective. I could not agree more.[1]

Next, you are undoubtably correct that the semantics of information, Shannon or otherwise, is observer-relative—e.g. given the same tree rings a knowledgeable observer learns more about the age of the tree than a naïve observer.

Onto your two objections.

Observer-relativity in information theory. After accepting that consciousness is observer-independent yet Shannon information is observer-relative, there are two distinct responses that retain the utility of information theory in studying consciousness.

1. It's sometimes said that Shannon theory gives us a beaker that determines the volume of a liquid but we remain ignorant of what the liquid consists of. This "volume" of semantic information, Shannon mutual information, is akin to syntax and is observer-independent. And as syntax is necessary for semantics, we can use mutual information to establish some observer-independent necessary conditions for consciousness.

2. We can tweak Shannon mutual information to make it observer-independent. Giulio Tononi does this by incorporating Judea Pearl's interventions into the foundation of his measures. This response implicitly assumes that, unlike Shannon mutual information, Pearl's causation (specifically, *information flow*[1] from a maximum entropy distribution) is

[1] Although it remains unclear to me that a science of consciousness *must* be observer-independent, I'd naturally prefer a science of consciousness that was.

> both epistemically-objective and observer-independent. Whether this is true or not I leave to the philosophers.[2]

Necessary conditions for consciousness. Just as simple behavioral experiments using verbal report establish sufficient conditions for consciousness, simple mathematical techniques can establish necessary conditions for consciousness. For example, in the photodiode thought experiment by counting the photodiode's internal states we know that it can have at most two different experiences. It can't have the separate experiences of red, green, and blue; the system's internal states just aren't there to permit it.

More sophisticated info-theoretic measures, Shannon or otherwise, can establish tighter necessary conditions for consciousness. For example, every state of awareness (non-reflective consciousness) conveys information (discriminates states) about the outside world. And without this information transmission, awareness is impossible. Therefore the system's "richness" or "magnitude" of awareness is upperbounded by the information its current internal state[3] conveys about the outside world.

Information theory can also quantify the irreducibility of the system's update rule.[3] You write,

> ...in experiencing a red square we "differentiate" the property of redness and the property of squareness, but the experience is "integrated" in that it "cannot be decomposed into the separate experience of red and the separate experience of a square."

We know the brain undergoes some physical process to "integrate" or "fuse" distinct percepts (e.g. redness and square) into a single experience (e.g. a red square). This physical process follows some algorithm. Therefore we can use algorithmic information theory to detect the presence of any process instantiated by the whole that is irreducible to the processes instantiated by the constituent parts "acting independently".[4]

[2]Tononi's measures based on Pearl's interventions are sometimes termed "non-Shannon" informations, but it's clearer to think of them as standard Shannon theory with the heart of Pearl. Genuine "non-Shannon" informations like Tsallis, Rényi, or Bar-Hillel/Carnap information are much more exotic creatures than information flow.

[3]Although it remains unclear precisely which among of the myriad of distinct notions of integration/irreducibility is most appropriate for consciousness, each one can be quantified.

[4]With some simplifying assumptions the space of "all possible algorithms" could likely be shrunk to the scope of algorithms/processes described by Shannon information theory.

Panpsychism in IIT. You write,

> Consciousness cannot be spread over the universe like a thin veneer of jam; there has to be a point where my consciousness ends and yours begins. For people who accept panpsychism, who attribute consciousness, as Koch does, to the iPhone, the question is: Why the iPhone? Why not each part of it? Each microprocessor? Why not each molecule? Why not the communication system of which the iPhone is a part? ... Consciousness comes in units and panpsychism cannot specify the units.

I can clarify some points here. The concern of "the units" is actually addressed by the original Balduzzi-Tononi 2008 paper[2]—particularly page 7 on complexes. Tononi asserts that only *complexes* possess phenomological experience. IIT states that a conscious entity S, by definition, is not subsumed by an entity T with strictly higher ϕ. Mathematically, if S has phenomological experience, then,

$$\phi(T) \leq \phi(S) \qquad \text{for } T \subset S \,. \tag{1}$$

Note this means there could still be a subset of S with higher phi—according to IIT, your whole brain remains conscious even if your prefrontal cortex has strictly higher ϕ. However, if your brain were embedded within an ultra-conscious hivemind, your brain would lose consciousness.

There are two oddities with this explanation. The first is that complexes can overlap, so the same unit can be part of two separate consciousnesses at once. The second is that, because the "greater than" property is a binary yes/no instead of a continuum, we can toggle human consciousness on/off with a single interconnect. Imagine we start hooking up human brains into a prototype hivemind. For simplicity, assume human each constituent human brain has exactly n bits of ϕ. While the hivemind has less than n bits of ϕ, each human brain is conscious. However, adding a final interconnect link causes the hivemind to exceed n bits of phi and requisitely renders all constituent brains unconscious. By adding and subtracting this final link we toggle the consciousness of the constituent brains. The scenarios above seem rather strange, but science has accepted stranger things, such as entangled particles transmitting information faster than the speed of light. Therefore I don't see these unusual properties as a deathblow to IIT.

A related criticism of IIT is that it remains hazy for how to properly account for systems operating at vastly different timescales. If we ourselves are conscious yet also embedded in an ultra-conscious galaxy at that integrates information on the timescale of millennia, it remains unclear how IIT would handle this case.[5]

Sincerely,
Virgil Griffith
virgil@caltech.edu

References

1. Ay D, Polani D. Information Flows in Causal Networks. Advances in Complex Systems 11(1) (2008) 17-41.

2. Balduzzi D, Tononi G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. PLoS Comput Biol 4(6) (2008).

3. Deweese M, Meister M. How to measure the information gained from one symbol. Network: Computation in Neural Systems 10(4) (1999) 325–40.

4. Tononi G. Integrated information theory of consciousness: an updated account. Arch Ital Biol 150(2-3) (2012) 56-90. doi: 10.4449/aib.v149i5.1388.

[5]There's a new paper[4] by Tononi that covers this, but it's sparse on the relevant equations.