The Integrated Information Theory: Important Insights but Not a Complete Theory of Consciousness

The Feelings of life Itself: Why Consciousness is Widespread but Can't be Computed

By Christof Koch, Cambridge Massachusetts, MIT Press, 2019, pp. 241.

Science has told us so much the universe around us. But can science explain the human mind itself? Can the eye that looks out at the cosmos turn inwards to look at itself? In *The Feeling of Life Itself*, Christof Koch makes an impassioned case not only that it can but that it has. The book is a deeply enjoyable journey into the mystery of consciousness, beginning with a whistle-stop tour of the history of the scientific study of consciousness, ultimately leading towards a wonderfully accessible introduction to the theory that Koch believes gives us the answers: Integrated Information Theory (IIT). Those like me who are passionate about this topic will learn a lot and enjoy the ride.

IIT has received a great deal of attention recently. It is not without controversy, with some arguing that it has absurd implications (Aaronson 2014) and some even arguing that it is meaningless (Pautz 2015). But this is perhaps to be expected from a scientific theory that proposes radical innovations to the very methods of science, and I'm inclined to think that radical innovations are going to be needed if we hope to explain consciousness. So overall, I'm enthusiastic about the project and I'm enthusiastic about this expression of it. But I do have one big disagreement with IIT as Koch and Giulio Tononi (the originator of IIT) present it. And so, in the spirit of friendly challenge, I will spend the rest of this piece articulating this disagreement.

Most neuroscientific theories are defended through purely empirical methods. Actually, that's not quite right, at least if 'empirical' is used to refer to research built on publicly observable data. The problem with the science of consciousness is that the phenomenon we're trying to study is essentially private: only you can observe your own experiences. A scientist can't look inside your head and see your feelings and sensations. Therefore, in studying consciousness, neuroscientists are forced to rely upon the testimony of test subjects regarding their private experiences, which are then correlated with activity in the brain that can be observed, using fMRI scanners or EEG. Nonetheless, there are rigorous and well-developed methods for working with this expansive conception of empirical enquiry.

The radical innovation of IIT is in its attempt to discern what consciousness is through introspective attention to consciousness itself. IIT proposes five 'axioms', truths about consciousness which, the theory claims, can be known merely through attending to one's own private experience. The next step is to bridge the gap from the mental to the physical by proposing five postulates, corresponding to the five axioms. Each postulate specifies a certain feature which, according to IIT, a physical system must satisfy in order to realise the feature of consciousness specifying by the corresponding axiom. So, for example, the first axiom is:

> Axiom 1: Any experience exists for itself; it is not dependent on anything outside of itself.

The corresponding postulate is:

> Postulate 1: For a system to exist for itself, it must have causal power over itself.

The theory then goes on to give a quite precise information-theoretic definition of what having causal power over oneself consists in, involving the capacity of a system to constrain its past and future states. All five axioms are summed up by Koch in the following:

> Every conscious experience exists for itself, it is structured, is the specific way it is, is one, and is definite. (p. 10)

In this way, by proposing the five axioms and translating them into postulates, the theory tries to identify information-theoretic criteria of consciousness somewhat independently of the standard methods of empirical testing.

One thing which is perhaps underexplored by proponents of IIT is the issue of whether we should think of the axioms as merely *necessary* for consciousness or both *necessary and sufficient* for consciousness. To be sure, even if the five axioms identify merely necessary features for consciousness, this is still potentially a crucial tool in homing in on the physical correlates of consciousness (assuming the translations from axioms to postulates hold up). But it leaves open the possibility that there are some physical systems that have all five features identified by the postulates but are still not conscious, because they lack certain *other* features which are also necessary for consciousness. It would also mean that we haven't completely accounted for consciousness, as we haven't identified which physical features are *enough* to ensure that a system is conscious. To take a specific example, many philosophers and scientists argue that consciousness essentially involves higher-order awareness, and some (Brown 2017) have criticized IIT for not including this feature of consciousness in its axioms.

However, let us suppose that the five characteristics do adequately capture necessary and sufficient conditions for consciousness. The next question is whether, in each case, the move from the axioms to the postulates is persuasive. Koch describes the move from axioms to postulates as a kind of *inference to the best explanation* (p. 75). This is, of course, a very common form of explanation in science; in chapter 2, Koch gives the example of the inference of the 19th century astronomer Le Verrier from irregularities in Uranus's orbit to the existence and location of an unknown planet, which turned out to be Neptune (p. 12). Koch also says that '[w]hat people mean by subjective feelings is precisely described by these five axioms' (p. 76). Thus, we can sum up the core thesis of IIT as follows:

> IIT-Core-Thesis: The features specified by the five postulates provide a complete explanation of all essential features of consciousness.

There is one big reason I don't believe the core thesis of IIT is entirely true. This is because it's inconsistent with a core thesis I defend in my book *Galileo's Error: Foundations for a New Science of Consciousness*. Like Koch, I think there are things we can know about the essential nature of consciousness via introspection. And one thing I think we can know is

that consciousness involves *qualities*: the blueness of a blue experience, the itchiness of an itch, the feel of cold ice. These conscious states are essentially defined by the qualities that characterize them. I argue at length in *Galileo's Error* (and at even greater length in my academic book *Consciousness and Fundamental Reality*) that you can't capture these qualities in a purely quantitative vocabulary. You can't convey in a purely quantitative language what it's like to see blue or to feel cold.

By a 'quantitative' vocabulary, I mean one that involves only mathematical and/or causal notions. It is pretty uncontentious that the vocabulary physical science uses to characterise matter is purely quantitative in this sense. Neurophysiological properties are defined either in terms of their causal role in the brain, or in terms of their chemical constituents. Those chemical constituents are defined either in terms of their causal relationships with other chemical entities, or in terms of their physical constituents. Ultimately, physical properties are characterised in terms of what they do, e.g. mass is characterised in terms of gravitational attraction and resistance to acceleration. The only concepts needed to describe all of this are causal and mathematical.

In other words, the core thesis of my book is:

> GE-Core-Thesis: Consciousness essentially involves qualities, and you can't capture the qualities of consciousness in a purely quantitative vocabulary.

Unfortunately, IIT-Core-Thesis and GE-Core-Thesis can't both be true, because the five postulates which are supposed to explain all features of consciousness are framed in a purely quantitative, information-theoretic vocabulary. We can perhaps make this clear by focusing on the third axiom, which would seem to encompass the qualitative aspect of consciousness. Koch describes it as follows:

> …any experience is highly *informative*, distinct because of the way it is. Each experience is informationally rich, containing a great deal of detail, a composition of specific phenomenal distinctions, bound together in specific ways. Each frame of every movie I ever saw or will see in the future is a distinct experience, each one a wealth of phenomenology of colors, shapes, lines, and textures at locations throughout the field of view. And then there are auditory, olfactory tactile, sexual, and other bodily experiences – each one distinct in its own way (p. 8).

The above refers to the qualities involved in various sense modalities. But when this is translated into the corresponding postulate, the informativeness of the experience is cashed out in terms of the capacity of a system to constrain its past and future states, which is the core component of IIT's information-theoretic framework.

According to GE-Core-Thesis, this cannot be done. If you could capture what it's like to see red in information-theoretic terms, then you'd be able to use that description to convey to a colour-blind neuroscientist what it's like to see red. But that's absurd; you can only fully grasp what it's like to see red when you actually have a red experience and attend to its qualitative character. No matter how much a colour blind scientist learns about the ways in which the brain states underlying our colour experience constrain their past and future

states, this will never give her a full understanding of what it's like to see those colours. Perhaps one can capture the *structure* of consciousness (the skeleton, as it were) in information-theoretic terms, but this kind of purely quantitative vocabulary simply doesn't have the resources to capture the qualitative character that fills out that structure (the meat on the bones).

Koch might object that the postulates do not need to *articulate* the qualities in order to *explain* them. But I can't see how you could explain the qualities in information-theoretic terms without being able to articulate them in information-theoretic terms. If I can't fully capture in information-theoretic terms what it's like see blue, then, for any state that's described in information-theoretic terms, we're going to be left with the question: 'But why does it feel like *this* [attends-to-the-blueness-of-the-experience] to be in *that* information-theoretic state?' This is a case where an *expressive limitation* (you can't articulate the qualities of experience using an information-theoretic vocabulary) entails an *explanatory limitation* (you can't explain the qualities of experience using an information-theoretic vocabulary).

One could, of course, just declare a brute identity between certain states characterized in information-theoretic terms and certain states characterized in qualitative terms. Many materialists do appeal to such brute identities. But at this point, I think one loses the right to claim to be *explaining* what is apparent to us on the basis of introspection, and this seems to me to undermine the distinctive introspection-based justification Tononi and Koch put forth in defence of IIT. If Koch wants to go for brute identities between mental states and physical states, then it seems to me that he should justify them on the basis of empirical findings alone. That is to say, he should offer a purely empirical argument that an identity between consciousness and maximal integrated information offers the best explanation of the empirical correlation we find between consciousness and maximal integrated information. He shouldn't claim to be *explaining* features of consciousness that are known on the basis of introspection, and then justifying IIT on this basis.

For these reasons, I don't believe IIT offers a fully adequate account of consciousness. The qualities of experience cannot be fully accounted for in information-theoretic terms. This doesn't by any stretch of the imagination mean that IIT has nothing to offer; but it does mean that it doesn't tell the full story. Essentially, I'm arguing that IIT doesn't have a fully adequate response to the 'hard problem of consciousness,' something which has previously been argued by Brown 2017 and Mindt 2017.

I suspect Koch might think these objections are overly philosophical. When considering the question 'Why should it feel like anything to be a maximum of integrated information?', Koch compares this question to the question of why the laws of quantum mechanics hold in our universe, or the question of why our universe is fine-tuned for life, saying:

> Speculations about ultimate "why" questions are enjoyable at the intellectual level. But they also contain more of a whiff of the absurd, trying to peek behind the curtains that hide the origin of creation only to find an endless set of further curtains. I will happily go to my grave knowing that in this universe, IIT characterizes the relationship between experiences and their physical substrate (p. 77).

Every theory has to start with some primitives, and there will always be the unanswerable question: 'Where did the primitives come from? Why is there something rather than nothing?' But that's not the kind of question we're dealing with here. IIT aspires to *explain* consciousness in terms of integrated information, and what is being pressed here is whether that explanation is adequate. Perhaps it would be appropriate for a substance dualist, i.e. someone who tries to explain consciousness in terms of immaterial souls, to reject the ultimate "why" question of why consciousness exists. The existence of immaterial souls is simply a basic and unexplained commitment of the dualist theory, and hence to ask a dualist 'Why are there immaterial souls?' would be akin to asking a physicalist 'Why is there matter?' But proponents of IIT don't just take human consciousness as an unexplained primitive; they want to explain it in terms of integrated information. My point is that if GE-Core-Thesis is true, that explanation fails.

Moreover, Koch and Tononi can hardly reject philosophical worries about their theory, as IIT is a deeply philosophical theory. It is defended not simply on the basis of empirical considerations, but on the basis of philosophical claims about what can be known on the basis of introspection, and how what can be known on the basis of introspection can be cashed out in information-theoretic terms. As such, it is reasonable to challenge it on philosophical grounds.

The problem of consciousness is radically different from any other scientific problem. Consciousness it not publicly observable, and its qualitative nature resists quantitative analysis. We need innovative approaches to deal with this, and IIT is a pioneering innovation which yields many important insights. For the reasons I have discussed, I don't believe that in its current form it constitutes a wholly adequate theory of consciousness. But the ideas of IIT may play a role in shaping future developments in the field. For example, as I discuss in my book, Hedda Hassel Mørch (2018) has developed the basic structure of IIT in the framework of Russellian panpsychism. We're living through exciting times for the science and philosophy of consciousness, and I can't wait to see what tomorrow will bring.

References

Aaronson, Scott (2014) 'Why I am not an integrated information theorist (or The Unconscious Expander),' *Shteltl-Optimiszed*. https://www.scottaaronson.com/blog/?p=1799.

Brown, Richard (2017a) 'Integrated information theory is not a theory of consciousness,' *The Usual Phlegm and Philosophy.* https://onemorebrown.com/2017/08/05/integrated-information-theory-is-not-a-theory-of-consciousness/

Brown, Richard (2017b) 'Integrated information theory doesn't address the hard problem,' *The Usual Phlegm and Philosophy.* https://onemorebrown.com/2017/08/13/integrated-information-theory-doesnt-address-the-hard-problem/

Goff, Philip (2017). *Consciousness and Fundamental Reality*, Oxford University Press.

Goff, Philip (2019) *Galileo's Error: Foundations for a New Science of Consciousness*, Rider/Pantheon.

Mindt, Garrett (2017). The problem with the 'information' in integrated information theory. *Journal of Consciousness Studies*, 24 (7-8), 130-154.

Mørch, Hedda Hassel (2018). 'Is the integrated information theory of consciousness compatible with Russellian panpsychism? *Erkenntnis* 84(5), 1065-1085.

Pautz, Adam (2015). 'What is integrated information theory a theory *of*?' https://philpapers.org/archive/PAUWII.pdf

Philip Goff, Assistant Professor of Philosophy

Department of Philosophy

Durham University

Durham, DH1 3HN

Email: Philip.a.goff@durham.ac.uk