Philosophy Now
a magazine of ideas

Philosophy Now – Issue 121
https://philosophynow.org/issues/121/The_Integrated_Information_Theory_of_Consciousness

The Integrated Information Theory of Consciousness

Hedda Hassel Mørch asks: what is IIT all about?

Consciousness is something with which we're all intimately familiar. It's the thing that goes away every night in deep sleep, and comes back when we wake up every morning, or whenever we start dreaming. It encompasses all our subjective feelings and experiences, ranging from the simple redness of red, to the complex depth of an emotion, to the ephemeral quality of thought. It's the one thing that is directly and immediately known to us, and it mediates our knowledge of the external world. This is how consciousness is defined by neuroscientist Giulio Tononi, the originator of the Integrated Information Theory of consciousness, or IIT for short. IIT is now one of the leading theories of consciousness in neuroscience.

According to IIT, consciousness is linked to *integrated information*, which can be represented by a precise mathematical quantity called Φ ('phi'). The human brain (or the part of it that supports our consciousness) has very high Φ, and is therefore highly conscious: it has highly complex and meaningful experiences. Systems with a low Φ, the theory goes, have a small amount of consciousness – they only have very simple and rudimentary experiences. Systems with zero Φ are not conscious at all.

IIT has radical implications. If IIT is true, we could in principle build a 'consciousness-meter' that tells us whether any system is conscious, and to what level: from comatose patients to infants; from simple animals and plants to robots and next generation AI. It also implies a kind of panpsychism, the view that all things are associated with some amount of consciousness [see the article by Philip Goff, Ed]. It would also have implications for the hard problem of consciousness: the philosophical question of why and how physical processes can give rise to subjective experience.

Information

Books, photographs and hard drives are typically regarded as containing a lot of information. But this information is mainly about other things: books describe events in the world, photographs depict external scenes, and so on. The information content also depends on human conventions about symbols and their meanings. In contrast, according to IIT, the only kind of information that matters for consciousness is the information a system has about itself. This information must be based on the system's causal powers, not on symbolic conventions.

To measure information of this kind, we ask: how much can we know about the previous and next state of the system by looking at the state of the system right now? For example, the current state of a typical human brain can tell us a lot about what that brain looked like a moment ago, and what it will look like in the next moment. There are a limited number of previous brain states that could possibly have caused its current state, and a limited number of future brain states that it could possibly cause. The brain is of course influenced by external conditions too, such as the sensory environment and bodily processes. But any such external conditions still leave a lot to be determined by the brain itself.

Compare this with another complex organ, the human retina. By looking at the current state of the retina, we learn a lot about what the environment in front of the retina was like a moment ago. We also learn about the next state of the visual processing system that takes input from the retina. But we don't learn much about the past and future states of the retina itself, because they are nearly completely fixed by the external environment – very little is left to be determined by the retina itself. This gives the retina very little information in IIT's sense.

How much information a system has about itself also depends on its number of possible states. A simple photodiode, that can be either on or off, can have very little information about itself, as its present state could rule out only one out of two possible states, at most. In contrast, the brain consists of billions of neurons, and there are endlessly many different combinations of neurons firing and not firing that are possible given most sensory, bodily and other background conditions. But knowledge of the current state of the brain rules out most of them: only a few of these combinations could have caused the current combination, and there are only a few combinations it in turn could cause. This gives the brain very high information about itself – IIT's first requirement for consciousness.

Integration

IIT's next requirement for consciousness is integration. Integration measures how much the information of a system depends on the interconnections between the system's parts. To determine it we ask: how much information is lost by cutting the system in two?

Consider a page of a book. The information in a book is mainly symbolic and about the external world, and therefore irrelevant for consciousness, but let's set this aside. If we tear the page horizontally in half, almost no information is lost. Reading one half page and then the other half page conveys the same information as reading the intact page. Therefore, the information on the page is not integrated. It's reducible to the sum of the information of the parts.

In the brain, in contrast (or more precisely, the areas relevant for our consciousness), every neuron is connected to thousands of other neurons, to form amazingly intricate structures. If the brain is cut in two, much of this structure would be lost, along with the information that depends on it. Any disconnected state will imply a very different past and future of the brain than an intact state would. This shows that the brain is a highly integrated system. Its information is not reducible to the sum of the information of its parts.

This is a key difference between brains and computers. A computer can have as much information as a brain – computers can have a similar number of possible states, and be at least as self-determining. But in a computer, at least as we make them today, every transistor is connected to only a few other transistors, so if we cut it in two much less structure would be lost. For this, and some further structural reasons (such as their modularity and feedforward connectivity), computers have very low integrated information, or Φ.

Maximality

Yet the fact that the brain has high integrated information does not fully explain its consciousness. IIT's third and final requirement is that a conscious system must be a *maximum* of integrated information. That is to say, it must have more integrated information than any overlapping system, including its own parts and any bigger system of which it itself forms part. This means, for example, that the area of the brain that directly supports our consciousness – the latest studies suggest some areas of the posterior cortex – must have higher Φ than any smaller neuron groups, individual neurons, molecules, and atoms that form part of it. It must also have higher Φ than the brain as a whole, the human body, human societies, the internet, and any other bigger system of which it forms a part, all the way up to the cosmos itself.

This claim has some interesting implications. If some smaller group of neurons within a larger brain area that normally supports consciousness suddenly became significantly more interconnected, and thereby surpassed the Φ of the larger area, then this smaller group would form its own consciousness separate from the larger whole. Or if the Φ of a normally conscious area suddenly dropped below the Φ of all smaller neuron groups at some level, its consciousness would dissolve into multiple lesser consciousnesses belonging to these neuron groups individually. Indeed, this could be what happens temporarily, in deep sleep: we think consciousness entirely disappears, but it might actually just change into a fragmented form (which is no longer recognizable as 'our' consciousness).

On the other hand, if the internet became more integrated than the human brain (when the internet is understood as a system that includes the brains of its users as parts, not just inputs to it) then the internet as a whole would become conscious and our own consciousnesses would be absorbed into it as parts! However, this would require that brains, computers and other elements of the internet became more closely interconnected than the neurons in our brains, so that physically speaking, the whole infrastructure would begin to look increasingly like an organism. It's safe to say that this is not on track to ever happening.

Third-Person Evidence

IIT tells a fascinating story about consciousness, but why should we believe it? Like any neuroscientific theory, IIT should mainly be judged by how well it explains the empirical data about consciousness.

One basic fact that we know is that human consciousness depends on the brain, and specifically, on some areas of the cerebrum, such as the posterior cortex. On the other hand, another part of the brain, the cerebellum, is important for motor functions, balance, and so on, but doesn't directly support consciousness. This poses a puzzle. The cerebellum contains more neurons than the cerebrum – 69 billion of the brain's total of 86 billion or so. So why is the cerebellum not more conscious than the cerebrum? IIT gives an answer: more neurons equals more information, but not more integration. A closer look at the cerebellum reveals that its neurons are far less interconnected than in the cerebrum. Therefore the cerebrum has much higher integrated information.

Another datum is that, contrary to what one might expect, the degree of consciousness doesn't correspond to the degree of brain activity. During epileptic seizures, brain activity increases dramatically, but consciousness disappears; and during deep, dreamless sleep, activity remains at normal levels. IIT explains this too. The patterns of activity seen during seizures and sleep are a highly regular series of bursts and silences, known as slow waves. These are patterns that can be shown to result from either low information or low integration.

IIT also makes new and testable predictions. By estimating, based on brain imaging, the Φ of patients who for various reasons (including strokes, brains lesions and anaesthetics) show no behavioral signs of consciousness, IIT can predict whether they are nevertheless conscious – either dreaming, or awake but 'locked in'. These predictions can be verified by comparing them with the results of other diagnostic tools, or sometimes the patients' own reports if they eventually wake up. So far, studies like this have corroborated IIT's predictions. The results are not conclusive though. There are rival theories of consciousness that emphasize the importance of, for example, fronto-parietal networks (a major one being the Global Workspace Theory developed by Stanislas Dehaene), and studies are often not precise enough to discriminate between them. Further experiments are needed to tell us more.

First-Person Evidence

Interestingly, Tononi did not come up with IIT purely by looking for patterns in third-person scientific data – from brain scans and so on. Rather, the theory was born from a philosophical argument based on phenomenology, which is first-person study of one's own consciousness. Tononi presents this as an essential

part of IIT's justification.

The argument starts from a list of five axioms – claims about consciousness that Tononi holds to be self-evidently true upon reflection on one's own consciousness. His first axiom holds that consciousness exists 'for itself', independently of external observers: it exists entirely for its own subject. The second axiom claims that consciousness is structured: it contains a variety of qualities at once; a mix of colors, sounds, emotions, thoughts, and so on (one might object that there are experiences of complete darkness that contain no qualities – but such an experience would still contain structure such as the left and right side of the empty visual field). The third axiom claims that consciousness is informative: like a painting, each experience specifies a 'scene' which is different from other possible 'scenes'. The fourth axiom holds that consciousness is integrated: the qualities within consciousness are unified under a single point of view, or we might say, by belonging to one and the same 'canvas'. Finally, the fifth axiom claims that consciousness is exclusive: the 'canvas' has an exact border, and any individual quality, such as a color or emotion, is either part of that canvas or not, never in between. Tononi holds that these axioms can be translated into a set of postulates that specify the physical counterparts of the properties they describe. These postulates are then given a mathematical interpretation, yielding the full version of IIT.

The physical counterparts to the axioms can be partially recognized in our earlier description of IIT. Because consciousness exists 'for itself', its physical counterpart must have information about itself. Because consciousness is structured, it must correspond to a complex physical structure. Because consciousness specifies one scene and thereby rules out others, the physical counterpart must rule out possibilities from a repertoire of possible physical states. Because consciousness is unified, its physical substrate must be physically integrated. Because consciousness is exclusive, conscious physical systems must have an exact physical border, defined by maximal Φ.

There are many questions one could raise about this argument. One might question whether the axioms are correct, or indeed whether there are any self-evident truths about consciousness at all. It can also be unclear how to precisely interpret the axioms, and what it means to translate them into the postulates about physical structure. Or one might object to the way they are translated into physical postulates, or the idea that it's even possible to do so.

Tononi's argument is nevertheless intriguing. It arguably stands to reason that first-person evidence should play an essential role in any theory of consciousness. After all, the first-person perspective is the only perspective from which consciousness can be directly observed. Consciousness can only be indirectly inferred from the third-person, external, perspective, from clues such as speech and behavior. Perhaps then the first-person perspective provides some crucial insight into the nature of consciousness. But even so, it remains to be seen whether IIT's particular first-person case can withstand closer scrutiny and criticism.

In sum, the combined empirical and philosophical evidence for IIT is controversial but significant. The evidence is far from conclusive, but it compares respectably to that of leading rivals, including Global Workspace Theory, predictive coding-based approaches, and quantum theories of consciousness, to mention a few. It has impressed several leading neuroscientists, including Christof Koch, one of the major pioneers of the field.

Artificial Intelligence & Consciousness

If IIT is correct, we could in principle measure the consciousness of any system by measuring its Φ. In

practice, Φ can't be precisely calculated except for very simple systems, because as complexity increases, the amount of computational power required to process the mathematical formulae involved approaches the impossible. But the Φ of most systems can nevertheless be roughly estimated by means of a variety of shortcuts and rules-of-thumb.

As mentioned, today's computers have very low Φ, for reasons including their sparse interconnectivity , regardless of how advanced they are. In the future, there might be computers and robots that equal or exceed humans in intelligence, understood in behavioral or computational terms; but as long as they're implemented by traditional hardware their Φ will remain insignificant, which means they will be either completely nonconscious or at best negligibly conscious. In other words, artificial intelligence does not necessarily imply artificial consciousness – that is, man-made systems with real subjective experience, as opposed to a mere outward simulation of consciousness.

Yet nothing in principle prevents computers from being made with integrated architecture. The limitations are rather practical: integrated systems are very difficult to design and engineer. Simply put, the more interconnections there are between the parts of a system, the easier it is to lose track of what's going on. The best way of engineering a highly integrated, and so conscious machine, may be by mimicking the structure of the brain – so-called 'neuromorphic architecture'; or alternatively, by mimicking the natural selection by which the human brain was created. It has been shown that integrated systems have some evolutionary advantages: in some ways, they are more efficient and more adaptable to change. By randomly adding new connections to a population of machines, and imposing conditions that select for the more efficient and adaptable ones, then repeating the process many times, one might succeed in selecting for integration, and by the same token, consciousness. Thus, there is a path to developing significant artificial consciousness, albeit a quite indirect and circumscribed one.

Animals & Plants

Another important question that's hard to answer without a theory is whether and to what extent animals are conscious – especially animals that are very unlike us, such as octopuses, fish, or insects. IIT implies that most animals probably are conscious. Most animal brains appear to be highly integrated. Going down the ladder of organic complexity, Φ, and consciousness, gradually decreases, but it never completely fades out. Even bacteria have a small amount of consciousness, because cells and organelles are integrated systems too. Plants, on the other hand, are probably not conscious, because individual plant cells can be estimated to have higher Φ than the plant as a whole – and consciousness requires maximal Φ. In terms of consciousness, then, a plant would be a society, not an individual.

Does this mean it's morally wrong to kill insects, fight bacteria, or destroy plant cells? The relationship between consciousness and moral status is intuitively a close one. If IIT is correct, one natural view is that moral status, just like consciousness, is a matter of degree. This would justify some common practices toward lower organisms – for example, the suffering and death of bacteria by penicillin is arguably worth the benefits it brings humans, given our vast difference in Φ. But it would still call for greater moral concern for most living organisms than we typically show.

The Inanimate World

This leads to another question, one that most of us normally wouldn't even think to ask: are inanimate objects

conscious? Just like current computers, chairs, rocks and most other macroscopic entities have negligible Φ – probably not enough to be maximal. But higher Φ might be found on some other scale for inanimate objects. Transistors, minerals and molecules, for example, all consist of mutually-interconnected smaller parts. They look like tiny integrated systems, possibly more integrated than any of the inanimate systems they compose. Further down, atoms consist of seemingly integrated sets of electrons, protons and neutrons. Even electrons, it has been argued, could have integrated structure, because physics no longer regards them as simple pointlike entities, but rather as complex fluctuations in fields. So, does consciousness go all the way down?

According to IIT, it probably does. Although it's not clear how exactly to apply the theory to fundamental physics, it's hard to avoid the interpretation that even particles have some Φ. The Φ of a particle would be vanishingly small compared to the brain. But as long as it's above zero, and is not surpassed by some greater system that it composes, such as a brain, particles must nevertheless enjoy some very basic form of subjective experience. IIT sets no minimal threshold of Φ required for consciousness.

The idea that even simple matter has some degree of consciousness is known as panpsychism. Panpsychism runs deeply counter to common sense, and many dismiss it as unscientific. Yet, Tononi openly stands by it insofar as it follows from IIT. After all, what is the evidence that particles are not conscious? That we have not observed them to be so is arguably irrelevant, because consciousness cannot be observed except in our own individual case. Furthermore, a long line of philosophers – from classics such as Gottfried Wilhelm Leibniz and William James, to contemporaries such as David Chalmers and Galen Strawson – have defended it. [For more about panpsychism, see Philip Goff's article in this issue.]

The Hard Problem of Consciousness

If it's correct, IIT solves what may be classified as one of the *easy* problems of consciousness, philosophically speaking: What sorts of physical states are essentially correlated with consciousness? The answer is: all and only those with maximal Φ. But there is also what is known as the *hard problem*: Why is consciousness correlated with any physical states at all? How does any physical state give rise to subjective experience?

Intuitively, it appears possible for any physical state to exist without being accompanied by subjective experience. This can be illustrated by the concept of a philosophical zombie, as introduced by David Chalmers in *The Conscious Mind* (1996). Philosophical zombies are physically identical to humans in every respect, including behavior, speech and internal neurological states, but have no subjective feelings and experience – there is nothing that it's like to be a philosophical zombie. Most of us have no problem imagining philosophical zombies, which suggests that we don't understand why they aren't possible. Now, consider Φ zombies – physical beings with maximal Φ, but no consciousness. It would seem that Φ zombies are just as conceivable as the other zombies, suggesting that we don't understand why maximal Φ must be accompanied by consciousness, either.

Yet IIT attempts to address the deeper, philosophically harder problem too, on the basis of its philosophical argument from phenomenological axioms to physical postulates. As discussed, IIT's philosophical argument is open to different interpretations and criticisms; but if first-person truths about consciousness can indeed be translated into physical postulates in a scientifically fruitful way, this implies a connection between the mental and the physical that's stronger than mere correlation. Tononi has described the connection as 'identity', but at the same time he explicitly holds that the first-person, experiential perspective on consciousness can never be replaced by any third-person, purely physical perspective. This indicates that the connection between the mental and physical is weaker than identity in the strict sense associated with

reductive materialism. If this in-between relation could be better understood, it might illuminate the hard problem.

Conclusion

Consciousness, according to IIT, is a matter of balance. On the one hand, it requires complexity and variation as conditions for high information. On the other, it requires unity and integration – the parts of a conscious system must be more strongly connected to each other than they are to anything else. IIT extracts these ideas from the first-person perspective, translates them into a precise mathematical measure, and tests the measure against third-person observations. So far, the results are promising, yet inconclusive. But if the theory does turn out to be on the right track, it has deep and radical implications for the place of consciousness in the natural order.

Hedda Hassel Mørch (pronounced 'Murk') is a post-doc in philosophy hosted by the NYU Center for Mind, Brain and Consciousness (co-directed by David Chalmers) and the Center for Sleep and Consciousness at University of Wisconsin-Madison (co-directed by Giulio Tononi).

• Resources for learning more about IIT (including a Φ calculator) can be found at integratedinformationtheory.org.