10-15-2022

Does integrated information theory make testable predictions about the role of silent neurons in consciousness?

Gary Bartlett

NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2022, **2022(1)**, 1–9
DOI: https://doi.org/10.1093/nc/niac015
Research Article

Does integrated information theory make testable predictions about the role of silent neurons in consciousness?

Gary Bartlett*,†

Department of Philosophy and Religious Studies, Central Washington University, 400 E. University Way, Ellensburg, WA 98926-7555, USA
†Gary Bartlett, http://orcid.org/0000-0003-0155-1843
*Correspondence address. Department of Philosophy and Religious Studies, Central Washington University, 400 E. University Way, Ellensburg, WA 98926-7555, USA. Tel: +1-509-963-2824; Fax: +1-509-963-1822; E-mail: gary.bartlett@cwu.edu

Abstract

Tononi *et al.* claim that their integrated information theory of consciousness makes testable predictions. This article discusses two of the more startling predictions, which follow from the theory's claim that conscious experiences are generated by inactive as well as active neurons. The first prediction is that a subject's conscious experience at a time can be affected by the disabling of neurons that were already inactive at that time. The second is that even if a subject's entire brain is "silent," meaning that all of its neurons are inactive (but not disabled), the subject can still have a conscious experience. A few authors have noted the implausibility of these predictions—which I call the disabling prediction and the silent brain prediction—but none have considered whether they are testable. In this article, I argue that they are not. In order to make this case, I first try to clarify the distinction between active, inactive (i.e. silent), and inactivated (i.e. disabled) neurons. With this clarification in place, I show that, even putting aside practical difficulties, it is impossible to set up a valid test of either the disabling prediction or the silent brain prediction. The conditions of the tests themselves are conditions under which a response from the subject could not reasonably be interpreted as evidence of consciousness or change in consciousness.

Keywords: consciousness; integrated information theory; predictions; testability; neural activity

Integrated information theory and the role of silent neurons

In a series of articles beginning in 2004, Giulio Tononi has maintained that, on his integrated information theory (IIT), consciousness is generated by inactive as well as active neurons.

In particular, Tononi has presented two striking predictions about the role of inactive, or "silent," neurons. The first is that a subject's conscious experience at a time can be affected by the disabling of neurons that were already inactive at that time. Call this the "disabling prediction." The second prediction takes the role of inactive neurons even further. It is that even if a subject's entire brain is "silent," meaning that all of its neurons are inactive (but not disabled), the subject can still have a conscious experience. Call this the "silent brain (SB) prediction."

Tononi *et al.* have increasingly emphasized the testability of IIT's predictions. As Tononi and Koch (2015) put it, "A theory is the more powerful the more it makes correct predictions that violate prior expectations" (p. 9; see also Oizumi *et al.* 2014; Tononi *et al.* 2016). Actually conducting tests, however, is not easy. Tononi (2017) says that so far, some simple predictions have been tested "only in an indirect and approximate manner, while others are in principle testable but technically demanding" (p. 251).

The disabling prediction and the SB prediction have drawn skepticism, but only for being counterintuitive (Fekete and Edelman 2011; Edelman and Fekete 2012; Klein 2019; Brette 2022; Pennartz 2022). Here, I consider whether they are testable.

As its name implies, IIT says that consciousness results from the integration of information. The relevant notion of information is causal and intrinsic. The more states a system S has (this is the "information"), the greater its capacity for consciousness; however, for that capacity to be realized, S's components must interact (the "integration"). If S is composed of elements e_1–e_n, then S contains integrated information to the extent that e_1–e_n influence each other. If (say) e_1's state is independent of those of e_2–e_n, then e_1 generates no information for S. However, if a state of e_1 constrains the past and future states of e_2–e_n (i.e. making some more likely and others less likely), then that state of e_1 generates information for S about its own past and future states.

Received: 27 April 2022; **Revised:** 9 August 2022; **Accepted:** 4 October 2022

Roughly, IIT identifies S's conscious state with the state that has the densest causal interdependence among its elements—what is called the "maximally irreducible cause–effect repertoire." The set of elements that specifies this repertoire is called a "main complex."

IIT is very abstract. Empirical theories of consciousness usually appeal directly to neural activity. IIT does not. Its framework of "axioms" and "postulates" is set up in purely phenomenological and mathematical terms. Canonically, the system elements are just logic gates.

It is also a vastly non-trivial task to apply IIT to a physical entity, such as the brain. One must specify the spatial extent, spatial grain, and temporal grain of the system, and the states of the elements. Ideally, these choices would be made via analysis of all possible decompositions of an entity, to identify the one with the maximally irreducible cause–effect repertoire. However, as Tononi *et al*. admit, such an analysis is possible only for toy networks (e.g. Oizumi *et al*. 2014). Concerning the brain, then, they have so far simply conjectured that the elements are neurons, and that qua elements, neurons have two states: active or inactive.

This conjecture then entails the disabling and SB predictions. In the IIT framework, active and inactive neurons can be equally relevant to the brain's cause–effect repertoire. To see why, it helps to think of a neuron's two states as simply α and β. Calling them "active" and "inactive," or even "on" and "off" (as Tononi *et al*. themselves often do), is misleading, as these names connote one state that "does something" and another that "does nothing." IIT denies that connotation. A neuron's inactivity is just as informative as its activity, for "[i]nactive elements of a complex specify a cause–effect repertoire (the probability of possible past and future states) just as much as active ones" (Tononi and Koch 2015, 10).

So, IIT in itself entails nothing about the role of silent neurons. It says nothing about neurons, and there is no such thing as a "silent" element. The two predictions arise only if silence is an informationally relevant neural state. This could turn out to be false. It could turn out that only active neural states specify a maximally irreducible cause–effect repertoire. So, it is more accurate to say that IIT permits the disabling and SB predictions, rather than directly entailing them. Nevertheless, permitting them is startling enough.

Crucially, IIT also says nothing about signals between elements. It thus rejects "the common assumption that neurons only contribute to consciousness if they are active in such a way that they 'signal' or 'broadcast' the information they represent" (Tononi and Koch 2015, 9). Consciousness is not the result of neurons signaling other neurons, as it is in most empirical theories, but rather of neural states occupying positions in a cause–effect state space.

IIT is thus agnostic about the processes or mechanisms by which the elements of a physical system affect each other. It is enough that they display causal interdependence. The cause–effect repertoire of a physical system is just a matter of the influence of each state of each element on the probability of the system's past and future states:

> Cause–effect power can be established by considering a *cause–effect space* with an axis for every possible state of a physical system in the past (causes) and future (effects). Within this space, it is enough to show that an "intervention" that sets the system in some initial state (cause), keeping the state of the elements outside the system fixed (background conditions), can lead with probability above chance to its present state; conversely, setting the system to its present state leads with probability above chance to some other state (effect). (Tononi 2017, 245)

In this way, IIT is similar to philosophical theories of consciousness, which are also characteristically silent on the operations of the physical substrate. As I will mention below, at least some of those theories entail the disabling prediction and perhaps the SB prediction too.

The disabling prediction and the silent brain prediction

Tononi (2004) offers a simple scenario to illustrate IIT's main implications. Imagine that you face a large screen. When turned on, it shows a homogeneous blue field. Activity in your brain's visual afferent pathways leads to the firing of blue-selective neuronal groups. You have been instructed to press a button when you see the blue field; so, the blue-selective activity in turn leads to the activation of motor pathways, which causes you to press the button. Meanwhile, many other neuronal groups, in the visual area and elsewhere, remain unaffected by the blue percept. Some are not firing; others are firing in order to serve a multitude of other functions.

Tononi says that IIT "makes several claims that lead to associated predictions" (Tononi 2004, 18) concerning this scenario. The most significant claim is that a neuron contributes to consciousness at a time if and only if it belongs to the main complex at that time. Therefore, blue-selective neurons are inside the main complex that specifies your conscious experience, whereas neurons in the afferent and efferent pathways, and those in many other parts of the brain, are not. However, the claim also entails that the activated, blue-selective neurons are not the only ones that contribute to your blue experience. As Tononi says, "the other groups of neurons within the main complex are essential to our conscious experience of blue even if, as in this example, they are not activated" (p. 19). The very inactivity of such neurons helps specify the informational structure of the main complex. This claim, in turn, entails the disabling prediction:

> Imagine that, starting from an intact main complex, we were to remove one element after another, except for the active, blue-selective one. If an inactive element contributing to "seeing red" were removed, blue would not be experienced as blue anymore, but as some less differentiated color, perhaps not unlike those experienced by certain dichromats. If further elements of the main complex were removed, including those contributing to shapes, to sounds, to thoughts and so forth, one would soon drop to such a low level of consciousness that "seeing blue" would become meaningless: the "feeling" (and meaning) of the quale "blue" would have been eroded down to nothing. (Tononi 2004, 19)

The prediction, in short, is that removing inactive neurons from the main complex will affect the subject's conscious experience at the very time of the removal.

The scenario has since been modified to involve temporarily "inactivating" inactive neurons rather than removing them (Tononi 2008, 2009; Balduzzi and Tononi 2009; Tononi and Koch 2015): hence my name, the "disabling" prediction. The difference between "inactivity" and "inactivation" is crucial. An inactive neuron still contributes to consciousness. In the neutral framing I suggested earlier, it is simply in state β rather than α but could

transition to α if needed. An inactivated (/disabled) neuron, by contrast, contributes nothing to consciousness because it generates no (intrinsic) information. That is because it is not responsive to the states of other elements.

Even more recently, Tononi has added that the disabling prediction applies to cases in which one simply disables the *connections between* the neurons:

> IIT predicts that changes in the efficacy of the connections among elements of the [physical substrate of consciousness] should lead to changes in experience even when these changes are not accompanied by changes in activity. A counterintuitive consequence of this prediction is that a brain area could contribute to an experience even if it is inactive but not if its connections or neurons are inactivated. Thus topographic visual areas would create visual space even in the absence of spiking activity but not if the horizontal connections within those areas are inactivated. Similarly, if the connections of neurons in colour areas are intact, the neurons would contribute to experience even if they are silent, by specifying negative colour concepts, such as when seeing a picture in black and white. However, if the connections are damaged, they would not specify any colour concepts, as with certain achromatopsic patients who do not even understand that the picture is missing colour. (Tononi *et al.* 2016, 459–60)

This variation on the disabling scenario accords with the core IIT principle that what matters to consciousness is the influence of elements on each other. One way to prevent that influence is to disable the elements themselves. However, another is to just cut them off from each other. Either way, the cause–effect repertoire of the system will be altered.

This variation emphasizes that for IIT, a system's counterfactual states—the states it could have occupied had different input been received—are crucial. In Tononi's original example, elements that would contribute to a red experience are removed while the subject is experiencing blue. Applying the most recent variation of the scenario to this example, the disabling prediction is that we can alter or cancel the blue experience just by disabling the connections of the "red" elements to the rest of the system. The system still moves through the same series of actual states because the "red" elements were not going to be called on to change their states anyway. However, the experience is nevertheless different than it would have been, just because of the fact that if those elements *had* been called on to change their states (e.g. if the screen had switched to red), they would not have done so.

Tononi *et al*. themselves emphasize how counterintuitive this prediction is (as in the above quote from Tononi *et al.* 2016). The SB prediction is even more counterintuitive. It posits consciousness in a brain in which no neurons at all are active, except at baseline.

This SB state must not be confused with comatose, vegetative, or minimally conscious states. These result from severe brain injury and thus involve widespread disabling of neural function (e.g. Laureys *et al.* 2004). By contrast, the SB state involves no injury at all.

The SB prediction says "that a brain where no neurons were activated, but were kept ready to respond in a differentiated manner to different perturbations, would be conscious (perhaps that nothing was going on)" (Tononi 2004, 19–20). As with the disabling prediction, the SB prediction has been regularly presented in subsequent publications, usually along with the idea that meditation might enable such a state of "naked awareness." For example:

> IIT predicts that, even if all the neurons in a main complex were inactive (or active at a low baseline rate), they would still generate consciousness as long as they are ready to respond to incoming spikes. An intriguing possibility is that a neurophysiological state of near-silence may be approximated through certain meditative practices that aim at reaching a state of "pure" awareness without content. (Oizumi *et al.* 2014, 17)

Other presentations of the SB prediction are found in Balduzzi and Tononi (2009), Tononi (2015, 2017), Tononi *et al.* (2016), and Tononi and Koch (2015).

In my view, part of what makes the disabling and SB predictions interesting is that they are not specific to IIT. I agree with Fekete and Edelman (2011) that the "problem of silent units," as they call it (i.e. the puzzle of how inactive parts of a system could contribute to its consciousness), is quite general. Any theory that pictures consciousness as an abstract property—not inherently biological or even physical—entails the disabling and SB predictions. Earlier, I noted IIT's similarity to philosophical theories of consciousness. Philosophers Maudlin (1989) and Antony (1994) criticized computationalist and functionalist theories of consciousness, respectively, for entailing the disabling prediction.[1] Tononi, to his credit, recognizes the entailment and forthrightly embraces its consequences.

Testability of the disabling and SB predictions: the general challenge

Notoriously, testing for consciousness is difficult because of the unobservability of the phenomenon. Behavioral responses to external stimuli are the standard measure. Neurophysiological responses are also used, typically via a prior correlation with a behavioral response. In both cases, the logic is that the observable response allows us to infer the occurrence of a conscious state that was the cause of (or in some neurophysiological cases, was identical to) the response. However, many responses can be caused by (or identical to) a non-conscious state rather than a conscious one. The challenge, then, is to distinguish responses that indicate a conscious state from ones that do not.

However, the disabling and SB predictions seem to pose an extra challenge. This is because the conditions of the test itself would seem to exclude the possibility of any response, and therefore the mere occurrence of a response would imply that the test itself was invalid.

The challenge is most apparent for the SB prediction. If a person's entire brain were silent, responses to stimuli would seem to be excluded *ex hypothesi*. Neurophysiological responses would surely show that the subject was not in the SB state; and if the subject were to respond behaviorally, one would assume that the movement was initiated by brain activity of some kind—meaning, again, that the SB state was not in effect.

As for the disabling prediction, the disablement is seemingly supposed to involve no change in brain activity. If so, then there would be no detectable neurophysiological response and there would be nothing to cause the subject to respond behaviorally. So, any response that did occur should lead us to conclude that the testing conditions were not met.

[1] For discussion of Maudlin, see Klein (2008), Bartlett (2012), and Klein (2019); and of Antony, see Bartlett (2014). It is less clear whether these theories entail the SB prediction; for related discussion, see Bartlett (2018). (In Bartlett (2012) I suggested that a scenario very similar to the disabling prediction could be used to test the thesis that consciousness supervenes on actual physical activity. I have changed my mind.)

I must immediately emphasize that these brief glosses are too quick. The testing conditions for the two predictions do not exclude all possible responses. Even the SB state does not involve total neurological immobilization, so the subject could in principle respond. However, I will argue that to do so, they must emerge from the SB state; so even if the response itself indicates a conscious cause, we cannot safely identify the SB state as having been that cause. So, the response could not be taken to show that the subject had been conscious while in the SB state. I will also argue that similar problems afflict the testing of the disabling prediction.

However, some subtle issues must be resolved in order to make these claims stick. In particular, we need to clarify what Tononi means by "silent" or "inactive" neurons.

Distinguishing active, inactive, and inactivated neurons

A "silent" or "inactive" neuron is not *completely* inactive. Unlike logic gates, neurons must undergo constant activity. To remain alive and functional, they must maintain metabolism and resting membrane potential. Tononi has increasingly acknowledged this fact. In Balduzzi and Tononi (2009) and Oizumi *et al.* (2014), the concept of inactivity is qualified by reference to a "baseline rate" of activation. More recently, the SB prediction has been framed in terms of the cortex being "almost" or "nearly" silent (Tononi 2015, 2017; Tononi *et al.* 2016). IIT now moots a set of "background conditions" in the brain, which are "[f]actors that enable consciousness, such as neuromodulators and external inputs that maintain adequate excitability" (Tononi *et al.* 2016, 452; see also Oizumi *et al.* 2014, esp. Suppl. Text 2).

What, then, does Tononi mean when he refers to "silent" or "inactive" neurons? What distinguishes an "active" neuron (which contributes to consciousness), an "inactive" neuron (which also contributes), and an "inactivated" neuron (which does not contribute)?

As I have noted, Tononi rejects the assumption that neurons contribute to consciousness only if they are signaling other neurons. I think that this is the key to the difference between activity, inactivity, and inactivation. Tononi often indicates that by calling a neuron "inactive," he means that it is not signaling or broadcasting to other neurons. For example:

> The assumption that neural elements that are active are broadcasting information often goes hand in hand with the corollary that inactive elements are essentially doing nothing, since they are not broadcasting anything. According to the IIT, this is not correct. (Balduzzi and Tononi 2009, 14)

> [The silent brain prediction] contrasts with the common assumption that neurons only contribute to consciousness if they are active in such a way that they "signal" or "broadcast" the information they "represent." (Tononi 2017, 252)

Similar remarks appear in Oizumi *et al.* (2014), Tononi (2015), Tononi and Koch (2015), and Tononi *et al.* (2016). The clear implication is that an active neuron is one that is signaling to other neurons. An inactive neuron would then be one that is *able* to signal but is not currently doing so; and an inactivated neuron, one that is unable to signal. The abstract IIT principle that an element contributes to consciousness only if it is influencing other elements would translate into the applied principle that a neuron contributes to consciousness only if it is able to signal other neurons. Note carefully: it is the *ability* to signal that constitutes influence, not the signaling itself. This is why inactive neurons carry influence, not just active ones.

But now: what exactly is meant by "signaling" or "broadcasting"? Clearly, an inactive neuron is not firing or spiking—i.e. generating action potentials—and it is commonly assumed that neurons communicate only via spikes. So, an obvious interpretation is that a signaling neuron is a spiking neuron, so that an inactive neuron is one that is not currently spiking but is ready to do so, and an inactivated neuron is one that is not ready to spike.

Here, we come to the key issue. Grant that an inactive neuron still has the baseline activity necessary to maintain spike readiness. This activity does not itself produce consciousness. However, such a neuron could also be engaged in a lot of other sub-threshold activity: more than just maintaining a membrane potential, yet not causing an action potential. Graded potentials are the obvious example, and such sub-threshold activity may be relevant to consciousness. Two pioneers of the neural correlates of consciousness program, one of whom is now Tononi's collaborator, more than 30 years ago wrote that the relevant neural activity might include "not only neurons that fire action potentials but also non-spiking neurons such as amacrine cells" (Crick and Koch 1990, 266). There are also theories (John 2001; McFadden 2020) on which consciousness depends on the electromagnetic fields generated by neural activity—not necessarily limited to spiking activity. In relation to these theories, it is worth noting that some recent studies (some featuring, again, Tononi's collaborator) suggest that neurons influence each other via their electrical fields (Anastassiou and Koch 2015; Faber and Pereda 2018). My point is not that these theories are correct, but that we cannot assume that spikes are the only activity that matters to consciousness. So, Tononi's advertised claim that inactive neurons contribute to consciousness might, on examination, become the far less counterintuitive claim that certain kinds of sub-threshold neural activity contribute to consciousness.

The issue boils down to this. Tononi might be making either of the following two claims about consciousness in the brain:

> Claim (1): Consciousness can be produced by the occurrence of non-spiking activity.
> Claim (2): Consciousness can be produced purely by the absence of spiking activity.

The difference is subtle, but it matters to the disabling and SB predictions. On (1), the two predictions are testable; but on (2), I shall argue, they are not. This is because on (1), the neural state that produces a conscious experience can also be the triggering cause of a subject's response to that experience. However, on (2), this is ruled out because any activity that might be involved is irrelevant to the production of consciousness. All that is relevant is that there is no spiking activity, for that is what makes it the case that it is in element state β ("off"/"inactive") rather than α.

So, which claim is Tononi making? I confess I am not completely sure. However, I think he should be making Claim (2), and there is evidence that this is what he and his colleagues intend.

First, IIT itself allows for (2), not just (1). So far as IIT itself is concerned, a neuron can be as inert as you like and still contribute to consciousness, so long as it remains able to transition to its other information-theoretic state(s). IIT certainly does not require, for example, that neurons must be engaged in graded potentials, or producing an electromagnetic field of a certain strength, in

order to contribute to conscious experience. So, it would be surprising if IIT's proponents were to limit themselves to (1). This limitation would largely strip the disabling and SB predictions of their counterintuitiveness and without motivation from IIT itself.

Second, if Tononi *et al*. actually have (1) in mind, it is odd that they never emphasize that the neurons they refer to as "inactive" are in fact active at sub-threshold levels. Indeed, one might expect them to not refer to them as "inactive" at all, as this would be misleading. As I have said, they acknowledge that these neurons must be active at baseline, but this is put forward as a mere enabling condition, not a component of consciousness itself.

Third, some things Tononi *et al*. say about the disabling prediction suggest that they are thinking of Claim (2). The disabling event entails the neurons in question switching from an inactive to an inactivated state. On Claim (1), that switch would be from a state of sub-threshold activation to a much lower-activation state. However, this is not how the disabling event is described. Recall that Tononi *et al.* (2016) say that one can merely disable the connections between the neurons; thus, there is no need to change the activation of the neurons themselves. Tononi and Koch (2015) also imply that there is no such activation change. They say that if the neurons were "pharmacologically or optogenetically inactivated, they would cease to contribute to consciousness [because] *even though their actual state is the same*, they would not specify a cause-effect repertoire" (p. 10, emphasis added). This suggests that the only change is to the neurons' role in a cause–effect network, not to their actual activity.

From here on, then, I will assume that Tononi *et al*. have in mind Claim (2). We now turn to the question of testing the SB prediction.

Testing the SB prediction?

In order to test the SB prediction, we will obviously need a subject who is actually in the SB state. This in itself may prove to be a major challenge.

Note further that it would not suffice for a subject to occupy the SB state only for a few milliseconds. Even if a response was reliably given whenever the state was momentarily attained, it would be impossible to determine with confidence whether it was caused by the SB state itself or by some active state that immediately preceded or succeeded it. Therefore, we would need the subject to remain in the SB state for at least several seconds and ideally for something like a minute.

The first problem, then, as Tononi has sometimes recognized (Balduzzi and Tononi 2009), is that it may not be possible for a brain to sustain a state in which all neurons are resting. Indeed, it may not be possible for a brain to enter the SB state at all; or perhaps, any brain that does so would not survive. If so, then the SB prediction is conclusively untestable.

For the sake of argument, however, let us assume that it is indeed possible for a brain to enter the SB state for a significant period and survive. Next is the question of how to induce the SB state. As noted earlier, Tononi *et al*. suggest that it might be achievable via deep meditation. Otherwise, an artificial process would be needed, perhaps akin to today's anesthetic procedures. Again, let us assume that some such procedure can be devised.

Finally: how we are to know for sure when the SB state has been fully achieved? This, too, is probably beyond our current capabilities. Let us assume, however, that at some future date, we will have the requisite neurophysiological measurement capabilities.

Suppose, then, that we have a subject who can either enter the SB state of their own accord or is willing to be induced into it. We would then arrange for the subject to signal their consciousness during the SB state. A simple discrimination task would do. Thus, we give the subject a button. We tell them that once their brain is silent, we will deliver one of two clearly distinct cues: say, a musical tone or the click of a ballpoint pen. They should press the button only when they hear (say) the click; they should ignore the tone. A test sequence might involve 20 trials. In 10 randomly selected trials we present the click; in the other 10, the tone. In each trial, the selected cue is presented only after the subject has maintained the SB state for at least 10 s. The cue is then presented at a randomly selected time within the next 10 s.

One might worry that it is unreasonable to expect a voluntary behavioral response from the subject. Such a worry would be groundless, however. While the SB state may superficially resemble a vegetative state, there is no (non-question-begging) reason to think that the former would impair the initiation of voluntary actions. Unlike a vegetative patient, a silent-brained subject has no brain damage. An impairment to voluntary action would presumably entail that some neurons are disabled, contrary to Tononi's own description of the SB scenario.

However, we now come to the key question: can we expect any response at all from a person whose brain is genuinely silent?

I do not think we can. The very conditions of the test make a valid positive response impossible. If the subject is genuinely in the SB state, they cannot make a response that would be evidence of consciousness.

Suppose that a particular test sequence produces the following results. On each of the 10 trials in which a click is presented, the subject presses the button within 1 s. On each of the 10 trials in which the tone (the distractor stimulus) is presented, the subject makes no response.

Suppose we say that this set of responses is evidence of consciousness. Then we are saying that the button presses were caused by the subject's having discriminated the two possible cues, and having decided that since a click was heard they should press the button. However, a response with this kind of causal origin entails not only a muscle contraction as a proximate cause but also a preceding activation in the somatic nervous system; and before that, activation in the motor cortex; and before that, activation of many other brain regions responsible for receiving the cue and deciding to respond to it.[2] None of this is compatible with the subject having an SB. *Ex hypothesi* we assume that on each trial, the subject is in the SB state right up until the presentation of the cue. Therefore, we should conclude that the cue raised the subject out of the SB state and that that was what caused the response—not the SB state itself. By analogy, you may wake a sleeping person by saying their name, but this does not show that they were conscious while asleep; only that there are processes that allow stimuli to rouse a sleeping person to consciousness. It is not thereby shown that the subject was already conscious before the cue was delivered. Similarly, our subject's responses do not show that they were already conscious while in the SB state.

Let me consider some objections.

Objection: the brain is not completely silent

The subject's brain is active at baseline—the neurons are maintaining a membrane potential. That activation could cause the button press.

[2] In fact, at least in the normal state of affairs, voluntary actions appear to arise not from an orderly, linear sequence of neural events but from a dynamic interaction between decision and action processes (Schurger and Uithol 2015).

Reply

It is true that the brain remains active in the SB state. The background conditions that enable consciousness remain in place. However, those conditions are defined as merely enabling consciousness, not producing it:

> [Background conditions] are the distal or proximal enabling factors that must be present for any conscious experience to occur—the heart must beat and supply the brain with oxygenated blood, various nuclei in the midbrain reticular formation and brainstem must be active, cholinergic release needs to occur within the cortico-thalamic complex, and so on. (Tononi and Koch 2015, 2)

This objection, therefore, says that the button presses in response to clicks are caused just by the background activation of the subject's brain—not by any activity that occurred subsequent to the cue. This is like saying that a burglar alarm was triggered just by the steady flow of electricity being delivered to the alarm as of its being turned on (say, 3 hours ago), not by the activity in the system that occurred after it detected motion. Background conditions are required simply to keep the brain alive and functional so that its cortical states can specify cause–effect repertoires that generate conscious experiences. Background conditions are (relatively) constant. Therefore, they cannot alone cause or explain an adventitious piece of behavior in response to a stimulus, because they do not vary with the receipt of the stimulus—just as the electrical power cannot alone explain the triggering of the burglar alarm.

Recall also from earlier that we are assuming that the important feature of inactive or silent neurons is the absence of spiking activity, not the presence of various kinds of non-spiking activity. So, there can be no implicit appeal here to, say, the role of graded potentials.

Objection: passive causation

The SB state could passively cause a response. Consider what Schaffer (2000) calls "causation by disconnection," in which an event occurs because of the disconnection or removal of something that was preventing it. As Schaffer explains, muscle contractions are caused in a way akin to the way a gun is caused to fire. In the gun, the pulling of the trigger releases a catch (the "sear") that was holding back the hammer. Similarly, in skeletal muscle, the arrival of a nerve signal removes a sheath that was preventing two sets of protein filaments from binding; and when those filaments bind, the muscle contracts. There could be many such causal "disconnections" in the chain of energy transfer between an action's cortical origin and the action itself. In particular, much causation in the brain will involve neurons that are *not* firing because an inhibitory neuron is preventing them from doing so. Of course, so far as we know, an inhibitory neuron must itself be firing in order to perform its function; so inhibition, as ordinarily conceived, cannot occur in an SB. However, what if there are actions that are in some way passively inhibited in the SB state, similarly to how the sheath passively inhibits muscle contractions? Then, the arrival of a cue might remove the inhibition, allowing the action to occur. Thus, an active response might be caused by a passive brain state.

Reply

It is sheer speculation that certain actions are passively inhibited in the SB state and that a specific cue (like a click) could release the inhibition. However, let us grant the premise. Furthermore, grant the metaphysical claim that the resulting response could be caused by the SB state. Even so, I contend that such a response could not plausibly count as the result of conscious volition, and thus, it could not plausibly count as evidence that the subject was conscious while in the SB state.

Suppose that the subject transitions out of the SB state on receipt of the cue and then presses the button. We then cannot rule out that the response was caused by the activity that followed the cue, rather than by the SB state itself, and so the response will not be evidence of consciousness in the SB state. Instead, it will more plausibly be evidence of consciousness after the subject has left the SB state, following receipt of the cue.

Suppose, on the other hand, that the subject remains in the SB state but still somehow presses the button. This means that there could be no processing of the cue in preparation for a response, and so it is not plausible that the subject *decided* to press the button. The response could only be an immediate and necessary consequence of the cue—akin to the firing of a gun or the contraction of a muscle, as in Schaffer's (2000) examples—and would thus appear to be more of an unconscious reflex than a conscious voluntary action.

The basic problem is that since one of two different responses is required (one of them being a passive or null response), and since the subject's brain must play a central causal role in determining which response is made, the subject's brain must transition into one of at least two different states as a causal precursor to making the appropriate response (i.e. pressing the button or the null response of inaction). However, their brain must therefore engage in some sort of adventitious activity in order to change its state—and so the subject cannot remain in the SB state.

Objection: neurophysiological measures

We should use a neurophysiological measure of consciousness instead of a behavioral one.

Reply

Even aside from the fact that neurophysiological measures are dependent on the prior establishment of a behavioral measure (cf. Irvine 2013), the only sign of consciousness we could possibly detect would simultaneously be a sign that the subject was no longer in the SB state.

Objection: dispense with cues

We should just train the subjects to recognize for themselves when they are in the SB state and to press the button once they reach it.

Reply

This does not escape the problem. Obviously, motor cortex activation is still necessary in order for the response to be delivered. There must also be some other prior activation corresponding to the subject's decision that they are now in the SB state and thus that they should now press the button. So, even if we set up the test so that the subject has to "self-cue" their response, this does not make a substantive difference. (It is also questionable whether a silent-brained subject would be able to self-cue a response.)

Objection: appeal to memory

The challenge here is surely not so different from the one faced in dream research, so surely it can be surmounted in the same way. We cannot ask a subject about their experience while they are dreaming, but we can still ask them immediately after they awake. Could we not do the same for a silent-brained subject?

Reply

This does not avoid the fundamental problem. The "delayed report" method works only if the subject, during the dream, encodes memories that they can later call up. However, encoding a memory requires the brain to undergo a change of some kind, so as to create the necessary neurological trace. Such traces can, of course, be created during dreaming, when the brain is very active. However, if the subject's brain is silent, then a trace cannot be created, for reasons like those given in my replies to the first two objections—the SB state rules out adventitious changes in the brain. So if the subject were to recall something afterward, it would only show that they had not consistently been in the SB state. (Remember, again, that we are assuming that sub-threshold activation is not relevant to consciousness; so, the IIT proponent cannot suggest that an experiential memory was created by such activation.)[3]

In sum, I find none of these objections compelling. I therefore maintain that the SB prediction is untestable, for the simple reason that the SB state itself makes a valid test impossible.

Testing the disabling prediction?

One might be more optimistic about testing the disabling prediction, given the less extreme neurophysiological conditions that are required. However, this optimism is misplaced.

To implement the testing procedure, we need a brain region—call it *R*—that is silent. Recall Tononi (2004, 19): "If an inactive element contributing to 'seeing red' were removed, blue would not be experienced as blue anymore." So, let us imagine that *R* is a region that contributes to the subject's visual experience of red; but they are seeing a homogeneous blue screen, so the neurons in *R* have only resting, or baseline, activation. Then, at a particular time *t*, we disable or inactivate *R*. The prediction is that at *t* the subject's conscious experience will change.

While perhaps less taxing than preparing a test of the SB prediction, preparing a test of the disabling prediction clearly still requires knowledge and techniques that we do not currently possess. We would need a way to precisely track the physiological state of well-defined brain regions in real time. We would also need a way to temporarily (and safely) disable a small brain region. For now, we can implement such deactivation only at a very gross level.[4] Tononi and Koch (2015) suggest that it might be achieved by pharmacological or optogenetic means. Cryostasis might also be suitable if the freezing is reversible.

Let us suppose that the methodological challenges can be overcome. The prediction says that at *t*, the moment of *R*'s disablement, the subject would undergo a change in their color experience. We would therefore instruct them to respond—perhaps, again, by pressing a button—if they notice any such change. As with the SB prediction, we could run a sequence of 20 trials. Each trial could last for, say, 60 s. In 10 randomly selected trials, we disable *R* at a randomly selected moment after the first 10 s; in the other 10 trials, we never disable *R*. Tononi's prediction is that the subject should reliably press the button within a second or two of each disabling event (or at the very least, that they should do so significantly more often than at times when the disabling event has not occurred).

The key question is whether the subject would be able to notice that the disabling event had occurred and thus report on it.[5]

Although the situation here is more complex than it was for the SB prediction, I think the same basic problem applies. *Ex hypothesi* region *R* is silent until *t*. Then, at *t*, it becomes disabled. Recall (from section "Distinguishing active, inactive, and inactivated neurons" above) that this disabling event ideally involves no change to activation in *R* itself. It just means that *R*, whatever it may be doing after *t*, no longer influences the state of other neurons. Now, as in the SB scenario, any response from the subject would entail muscle contractions, which in turn would require a preceding activation in the somatic nervous system and (before that) the motor cortex. In much the same way that it seems impossible for such a cascade to be initiated by the SB state, it seems impossible for it to be initiated by the disabling of *R*. The only way that the disabling event could trigger the necessary cascade is if, before *t*, *R* had been in some way affecting (some part of) the rest of the brain so that that effect was abruptly cut off at *t*. However, *R*'s having such an effect before *t* is inconsistent with its inactive state.

To be sure, according to IIT, *R* was exerting an influence over the rest of the brain before *t* due to its readiness to signal other neurons—which means it was in informational state β rather than α, to use my neutral terminology. That informational state was exerting a constraint over the past and future states of the system. However, we are here concerned with neurophysiology, not information, and at the neurophysiological level, there is no explanation for why the disabling of *R* should cause the subject to respond at all.

Strikingly, Tononi himself has indicated that the disabling event would have no effect on a subject's behavior. In a recent article, he presents the disabling prediction as follows:

> IIT predicts that a particular brain area can contribute to experience even if it is inactive, but not if it is inactivated. For example, if one were presented with a plate of spinach drained of color, green-selective neurons in the color areas would remain inactive. Thus, one would experience and report strange spinach that is gray rather than green. By contrast, if the same area were not just inactive, but inactivated due to a local lesion, the phenomenal distinctions corresponding to colors would be lacking altogether. While presumably one would still report that the spinach is "gray," in this case "gray" cannot mean the same as when color areas are intact, i.e. not green, not red, and so on. (Tononi 2015)

The subject's color experience would change: "'gray' cannot mean the same as when color areas are intact" (compare Tononi 2004, as I quoted earlier: "blue would not be experienced as blue anymore"). One would then expect the subject's report to reflect that change. While they may find their new experience hard to describe ("It's colored but not colored… It's like the space behind my head!"), they would not just continue to say, "It's gray." Yet, Tononi appears to state that indeed this is exactly what they would

[3] What if the memory encoding mechanisms are not part of the main complex but are in some other part of the brain that is not inactivated? (Thanks to a reviewer for this suggestion.) In response, first, I am not aware of any direct reason to think that the memory mechanism might be excluded from the main complex, so the hypothesis feels a bit *ad hoc*. Second, and more importantly, even if we accept the hypothesis, it is unclear how the memory mechanism could record anything if there are no relevant changes in the part of the brain it is monitoring.

[4] In the Wada test (see, e.g. Abou-Khalil 2007), one brain hemisphere is anesthetized by injecting sodium amobarbital via the internal carotid artery. Such an intervention is not suited for testing the disabling prediction, as it affects much too large a swath of the brain, a lot of which would have been active immediately prior to anesthesia.

[5] One might worry that the change may be too subtle to be noticed and reported on. However, this worry seems unmotivated. There is no (non-question-begging) reason to think that the disabling event we are imagining has to be subtle. Indeed, the way Tononi himself has described it makes it sound like it should be very obvious.

do. He writes that the subject would "still report that the spinach is 'gray"' after the green-selective neurons were disabled.

If what Tononi himself says here is right, then the disabling prediction is untestable. Despite the change in the subject's experience, there can be no change in their behavioral response—because (I have argued) there is no relevant change in the activity in their brain.

Nevertheless, defenders of IIT may advance some of the same objections they offered against my argument concerning the SB prediction. Let me briefly address these.

Objection: the brain is not completely silent

Again, one might try to argue that *R*'s baseline activation could have caused the response.

Reply

Again, this is fruitless. Baseline activation only preserves metabolism and functionality. It cannot be the cause of a discrete piece of behavior, and we have already ruled out other varieties of sub-threshold activation (see the section "Distinguishing active, inactive, and inactivated neurons").

Objection: passive causation

Perhaps *R*'s baseline activation inhibits the activation of other brain regions so that, when *R* is disabled, the inhibition is released, triggering a cascade of activation that leads to the subject pressing the button.

Reply

It is not clear that this is compatible with the stipulation that the disabling event involves no change in *R* itself. For example, if *R* does not change, how could it go from inhibiting some other event to not inhibiting that event? However, let us accept that the scenario is possible, albeit unlikely. Still, as with the passive causation objection in the SB case, a response generated in this way could not plausibly indicate a conscious volition.

Objection: appeal to memory

We should have the subject recall any changes in their experience after the test is completed, rather than expecting them to respond at the time.

Reply

A delayed response is still just as problematic as an immediate one. Suppose the subject does indeed recall a change in their experience at about the time *R* was disabled, and suppose we are confident that this report reflects an actual change in their experience at that time. We still must ask: what enabled the subject to make this report? The only reasonable answer is that some neurological trace—a memory—of the change was created at the time. However, how did that happen? We now face the same basic problem as before: if region *R* had been genuinely silent, then its disabling at *t* could not have caused the creation of a memory trace. The only way this could have occurred is if *R* was not silent after all, thus invalidating the test.

Concluding observations

The disabling prediction and the SB prediction are not testable. They cannot provide evidence that silent neurons can contribute to consciousness.

Of course, this is not to say that silent neurons cannot contribute to consciousness. For all I have argued, it remains possible that they do. It remains possible, that is, that disabling some inactive neurons can change a person's conscious experience, and even that a brain whose neurons are all inactive might still support conscious experience.

Interestingly, in his 2015 article—the same one, ironically, in which I think he effectively admits that the disabling prediction is untestable—Tononi distinguishes between "predictions" of IIT and "extrapolations" of IIT. In the latter category, he seems to put claims that are not testable, such as the claim that "a simple but large two-dimensional grid of appropriate physical elements could be highly conscious, even if it were doing 'nothing' (all binary elements off), and even if it were disconnected from the rest of the world." It is surprising that, having made this distinction, he still categorizes the claims about the role of silent neurons as predictions. It seems to me that they would be much better categorized as extrapolations.

As I noted earlier, IIT is not alone in these commitments regarding silent neurons. Some may say that a theory that implies that neurons can contribute to consciousness while doing nothing thereby provides its own *reductio*. However, it is unclear how much weight we should give to intuition here. Schwitzgebel (2014) argues that the correct metaphysics of the mind will inevitably be contrary to common sense. Despite the weirdness of IIT's claim about the role of inactive neurons, it remains possible that, in the end, we will have to accept it.

However, not, I think, just yet. In closing, let me point out a curious aspect of Tononi's position which might give us pause.

I have noted that Tononi *et al.* have often speculated that the SB state might be a state of "pure" or "naked" conscious awareness, devoid of content. However, I do not think that IIT itself supports this speculation. In suggesting it, Tononi *et al.* reveal an implicit commitment to a more significant role for neural activation than their theory allows.

The common intuition is that if one's brain is doing nothing, then one's mind would also be doing nothing. However, IIT's proponents do not see it this way. This is because of their belief that an inactive neuron can carry as much information as an active one; so that it can contribute as much to the subject's conscious experience as an active one.

So far, so good—albeit counterintuitive. However, Tononi *et al.* then speculate that the conscious experience of the SB state would be one of nothingness. This is surprising. From an informational perspective, the SB state is simply one among the vast number of states that a brain can occupy. *A priori*, there is no reason to have any particular expectation about the phenomenology of that state—any more than there would be for any other state we might select. Why could the SB state not produce, say, an experience of deep happiness? Or of a full-body itch? Or of warmth? Or of nausea?

I think that Tononi *et al.* are tacitly yielding to the pull of the common intuition noted above. Even though they claim that a silent-brained subject would be conscious, their speculation that the subject would nonetheless have a null or empty phenomenology is a symptom of the pull of the intuition. IIT itself provides no reason for that speculation.

Also telling is the IIT theorists' frequent suggestion that the SB state might be achieved through meditation. Partly, no doubt, they make this suggestion because of the expectation that the SB state would be a state of pure awareness.[6] Yet, there is no evidence

[6] Many forms of meditation actually involve focusing one's awareness on a particular thing, such as the breath. A state of pure awareness is the goal only of a few varieties, such as Daoist apophatic meditation (e.g. Roth 2015).

that a state of pure awareness correlates with anything like the SB state.[7]

Perhaps also contributing to the apparent connection between meditation and the SB state is the fact that meditation characteristically involves motionlessness, combined with the intuition that a silent-brained subject would have to be motionless. I have argued, of course, that this intuition is correct and that it makes a test of the SB prediction impossible. One might therefore expect Tononi to resist this intuition, as doing so would more easily allow for responses from a silent-brained subject. Yet, his association of the SB state with meditation suggests that he assumes that a silent-brained subject would indeed be inert. Again, I see nothing in IIT itself which requires this assumption. If, e.g. the SB state were to produce extreme itchiness (and again, IIT offers no reason why it could not), then a silent-brained subject might be expected to be constantly scratching.

It might now be pointed out that what I have just said is exactly why we should think that the SB state must be associated with an empty phenomenology. For how could a silent-brained subject possibly be scratching? And if they cannot possibly scratch, then how can they possibly feel an itch—or, indeed, anything? Isn't pure awareness without content the only phenomenology that is compatible with the subject's serene unresponsiveness?

However, this line of thought is at least as good an argument for the conclusion that a silent-brained subject would be unconscious as it is for the conclusion that they would experience pure awareness. My point here is that there is no reason to associate the SB state with a meditative state; and the fact that Tononi and colleagues make that association betrays their own tacit sense that there is, after all, something unusual about the SB state that is not captured by IIT itself.

Acknowledgements

Very great thanks to two anonymous reviewers for this journal, who offered many suggestions that improved the paper immensely. Thanks also to an audience at the Southern Society for Philosophy and Psychology.

Conflict of interest statement

None declared.

References

Abou-Khalil B. Methods for determination of language dominance: the Wada test and proposed noninvasive alternatives. *Curr Neurol Neurosci Rep* 2007;**7**:483–90.

Anastassiou CA, Koch C. Ephaptic coupling to endogenous field activity: why bother? *Curr Opin Neurobiol* 2015;**31**:95–103.

Antony MV. Against functionalist theories of consciousness. *Mind Lang* 1994;**9**:105–23.

Balduzzi D, Tononi G. Qualia: the geometry of integrated information. *PLoS Comput Biol* 2009;**5**:e1000462.

Bartlett G. Computational theories of experience: between a rock and a hard place. *Erkenntnis* 2012;**76**:195–209.

Bartlett G. Against the necessity of functional roles for conscious experience: reviving and revising a neglected argument. *J Consc Stud* 2014;**21**:33–53.

[7] There is evidence that meditation can indeed change one's brain activity (e.g. Panda *et al.* 2016; Vivot *et al.* 2020), but the changes that are observed do not involve a universal *decrease* in activity.

Bartlett G. Functionalism and the problem of occurrent states. *Philos Quart* 2018;**68**:1–20.

Brette R. Does the present moment depend on the moments not lived? *Behav Brain Sci* 2022;**45**:19–20.

Crick F, Koch C. Towards a neurobiological theory of consciousness. *Semin Neurol* 1990;**2**:263–75.

Edelman S, Fekete T. Being in time. In: Edelman S, Fekete T, Zach N (eds.), *Being in Time: Dynamical Models of Phenomenal Experience*. Amsterdam/Philadelphia: John Benjamins, 2012, 81–93.

Faber DS, Pereda AE. Two forms of electrical transmission between neurons. *Front Mol Neurosci* 2018;**11**:427.

Fekete T, Edelman S. Towards a computational theory of experience. *Conscious Cogn* 2011;**20**:807–27.

Irvine E. Measures of consciousness. *Philos Compass* 2013;**8**:285–97.

John ER. A field theory of consciousness. *Conscious Cogn* 2001;**10**:184–213.

Klein C. Dispositional implementation solves the superfluous structure problem. *Synthese* 2008;**165**:141–53.

Klein C. Computation, consciousness, and "Computation and consciousness". In: Sprevak M, Colombo M (eds.), *The Routledge Handbook of the Computational Mind*. London/New York: Routledge, 2019, 297–309.

Laureys S, Owen AM, Schiff ND. Brain function in coma, vegetative state, and related disorders. *Lancet Neurol* 2004;**3**:537–46.

Maudlin T. Computation and consciousness. *J Philos* 1989;**86**:407–32.

McFadden J. Integrating information in the brain's EM field: the cemi field theory of consciousness. *Neurosci Conscious* 2020;**6**:niaa016.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 2014;**10**:e1003588.

Panda R, Bharath RD, Upadhyay N *et al.* Temporal dynamics of the default mode network characterize meditation-induced alterations in consciousness. *Front Hum Neurosci* 2016;**10**:1–12.

Pennartz CMA. What is exactly the problem with panpsychism? *Behav Brain Sci* 2022;**45**:39–40.

Roth HD. Daoist apophatic meditation: selections from the classical Daoist textual corpus. In: Kojmathy L (ed.), *Contemplative Literature: A Comparative Sourcebook on Meditation and Contemplative Prayer*. Albany: SUNY Press, 2015, 89–144.

Schaffer J. Causation by disconnection. *Philos Sci* 2000;**67**:285–300.

Schurger A, Uithol S. Nowhere and everywhere: the causal origin of voluntary action. *Rev Philos Psychol* 2015;**6**:761–78.

Schwitzgebel E. The crazyist metaphysics of mind. *Australas J Philos* 2014;**92**:665–82.

Tononi G. An information integration theory of consciousness. *BMC Neurosci* 2004;**5**:1–22.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 2008;**215**:216–42.

Tononi G. An integrated information theory of consciousness. In: Banks WP (ed.), *Encyclopedia of Consciousness*. Oxford/San Diego: Academic Press, 2009, 403–16.

Tononi G. Integrated information theory. *Scholarpedia* 2015;**10**:4164.

Tononi G. The integrated information theory of consciousness: an outline. In: Schneider S, Velmans M (eds.), *The Blackwell Companion to Consciousness*. 2nd edn West Sussex: Wiley, 2017, 243–56.

Tononi G, Boly M, Massimini M *et al.* Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.

Tononi G, Koch C. Consciousness: here, there and everywhere? *Philos Trans R Soc London B* 2015;**370**:1668.

Vivot RM, Pallavicini C, Zamberlan F *et al.* Meditation increases the entropy of brain oscillatory activity. *Neuroscience* 2020;**431**:40–51.