

5-9-2019

# The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness

Adrian Doerig

*EPFL – École polytechnique fédérale de Lausanne*

Aaron Schurger

*Chapman University, schurger@chapman.edu*


Kathryn Hess

*EPFL – École polytechnique fédérale de Lausanne*

Michael H. Herzog

*EPFL – École polytechnique fédérale de Lausanne*

Follow this and additional works at: [https://digitalcommons.chapman.edu/psychology\\_articles](https://digitalcommons.chapman.edu/psychology_articles)

 Part of the [Cognition and Perception Commons](#), [Cognitive Psychology Commons](#), [Other Psychiatry and Psychology Commons](#), [Other Psychology Commons](#), [Philosophy of Mind Commons](#), and the [Psychological Phenomena and Processes Commons](#)

---

## Recommended Citation

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, 72, 49-59. <https://doi.org/10.1016/j.concog.2019.04.002>

This Article is brought to you for free and open access by the Psychology at Chapman University Digital Commons. It has been accepted for inclusion in Psychology Faculty Articles and Research by an authorized administrator of Chapman University Digital Commons. For more information, please contact [laughtin@chapman.edu](mailto:laughtin@chapman.edu).

---

# The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness

## Comments

This article was originally published in *Consciousness and Cognition*, volume 72, in 2019. DOI: [10.1016/j.concog.2019.04.002](https://doi.org/10.1016/j.concog.2019.04.002)

## Creative Commons License

Creative

Commons

This work is licensed under a [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/).

License

## Copyright

The authors



# The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness

Adrien Doerig<sup>a,\*</sup>, Aaron Schurger<sup>b,c,d,e</sup>, Kathryn Hess<sup>f</sup>, Michael H. Herzog<sup>a</sup>

<sup>a</sup> Laboratory of Psychophysics, Brain Mind Institute, EPFL, Switzerland

<sup>b</sup> INSERM, Cognitive Neuroimaging Unit, NeuroSpin Center, Gif sur Yvette 91191, France

<sup>c</sup> Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin Center, Gif sur Yvette 91191, France

<sup>d</sup> Department of Psychology, Crean College of Health and Behavioral Sciences, Chapman University, Orange, CA, USA

<sup>e</sup> Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Irvine, CA, USA

<sup>f</sup> Laboratory for Topology and Neuroscience, Brain Mind Institute, EPFL, Switzerland

## ARTICLE INFO

### Keywords:

Consciousness  
Theories  
Causal structure  
IIT  
RPT  
Neural networks

## ABSTRACT

How can we explain consciousness? This question has become a vibrant topic of neuroscience research in recent decades. A large body of empirical results has been accumulated, and many theories have been proposed. Certain theories suggest that consciousness should be explained in terms of brain *functions*, such as accessing information in a global workspace, applying higher order to lower order representations, or predictive coding. These functions could be realized by a variety of patterns of brain connectivity. Other theories, such as Information Integration Theory (IIT) and Recurrent Processing Theory (RPT), identify *causal structure* with consciousness. For example, according to these theories, feedforward systems are never conscious, and feedback systems always are. Here, using theorems from the theory of computation, we show that causal structure theories are either false or outside the realm of science.

## 1. Causal structure theories of consciousness

### 1.1. Consciousness and empirical data

We wake up every day and transition from an unconscious to a conscious state. Surely, there is something to explain. In binocular rivalry and visual masking, we can render clearly visible stimuli invisible. Surely, there is something to explain here too. These examples and many others are routinely used by the scientific community as a means to study consciousness and are at the heart of all empirically-minded theories of consciousness (Fig. 1a). Because of the subjectivity of consciousness, the dependent measures in these experiments are subjective reports (or other measurements known to reliably correlate with subjective reports). We cannot use measures of brain activity as a-priori indicators of consciousness because we want to understand how brain activity gives rise to consciousness in the first place.

### 1.2. Causal structure theories of consciousness

Theories of consciousness aim to explain how changes in the proposed mechanism lead to changes from unconscious to conscious

\* Corresponding author.

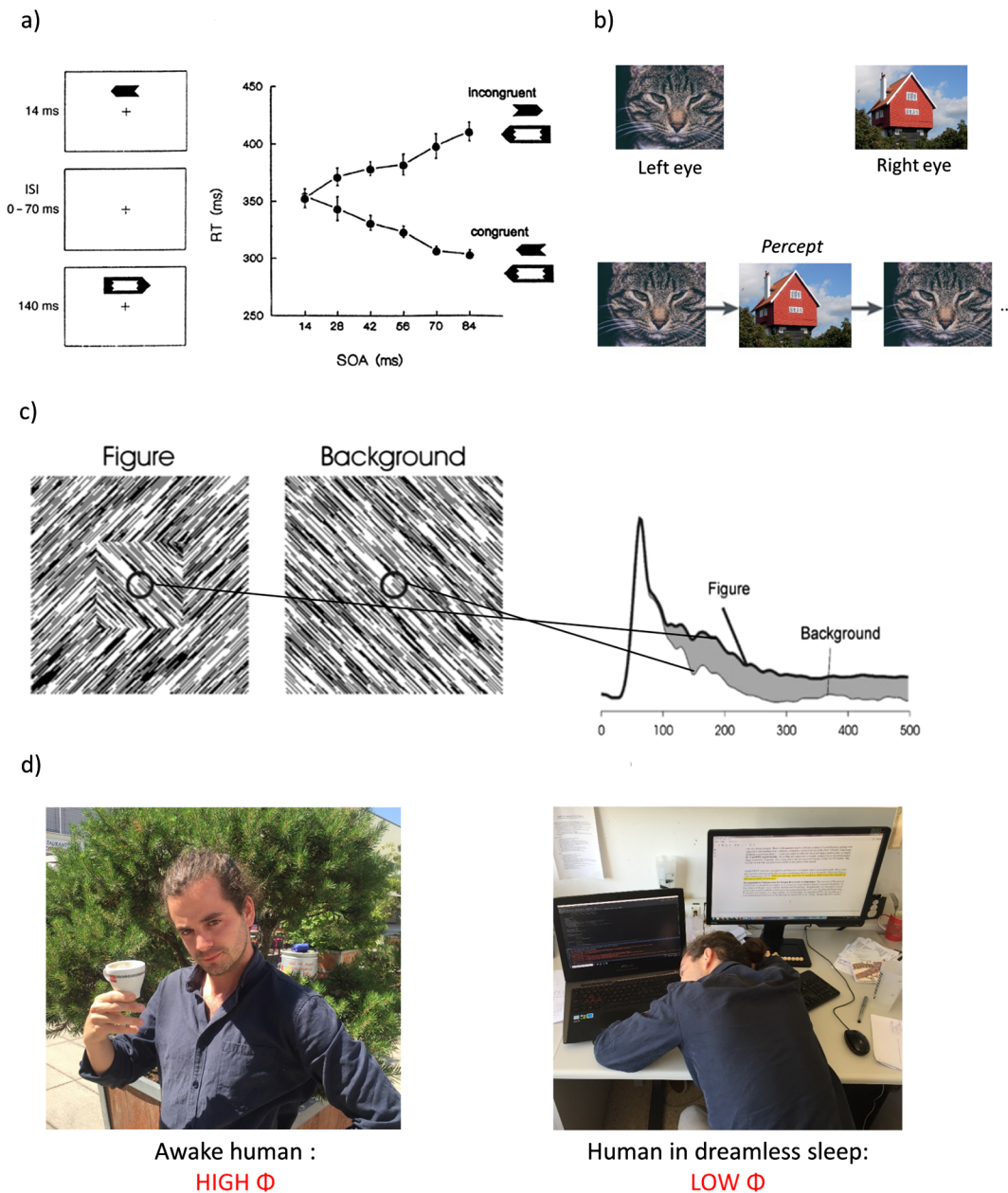
E-mail address: [adrien.doerig@gmail.com](mailto:adrien.doerig@gmail.com) (A. Doerig).

<https://doi.org/10.1016/j.concog.2019.04.002>

Received 20 November 2018; Received in revised form 30 March 2019; Accepted 1 April 2019

Available online 09 May 2019

1053-8100/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



**Fig. 1.** Empirical experiments to investigate consciousness. (a) Visual masking. A briefly presented prime (left pointing arrow) can be rendered invisible by a trailing mask (arrow with central gap), depending on the time interval between them (Inter-Stimulus Interval, ISI). Participants are asked whether the clearly visible masking arrow points either to the left or the right. Going from long to short ISIs between target and mask leads to a change from full target visibility to zero visibility. When the prime and the mask arrows point in the same direction (congruent trials), reactions are faster than when the arrows point in opposite directions even when the prime is unconscious (incongruent trials; SOA = ISI + prime duration, 14 ms here). Empirical theories of consciousness aim to show that their proposed mechanism explains when and why the target is (in)visible. Reproduced with permission from [Vorberg, Mattler, Heinecke, Schmidt, and Schwarzbach \(2003\)](#). Copyright (2003) National Academy of Sciences, U.S.A. (b) Binocular rivalry. Different images are shown to the left and right eye. In this example, a cat face is presented to the left eye and a red house to the right eye. When the images are not compatible, only one single image is consciously perceived at a time. After a few seconds, there is a switch and the other image is perceived. For each image, there are conscious (it is perceived) and unconscious (the other image is perceived instead) alternatives. Empirical theories of consciousness need to provide a mechanism to explain when and why each image is (in)visible. (c) RPT & Visual masking/anaesthesia. On the left, a central figure is consciously perceived. This is reflected by elevated neural activity after 100 ms, shown on the right. The elevation is mediated by recurrent connections from higher to lower areas of the visual system. When no central figure is consciously perceived, because only the background is presented, or during anaesthesia, the elevated activity is not present. (d) IIT & Wakefulness vs. Dreamless sleep. Consciousness changes from wakefulness to dreamless sleep. IIT proposes to explain this change by changes in causal structure (quantified by a number,  $\Phi$ , see main text). In accordance with IIT, experiments have shown that a practical proxy of  $\Phi$  is high during wakefulness, and low during coma or slow-wave sleep. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

states and vice versa. To do so, a number of theories propose that the essential element for understanding consciousness is *how* parts of a system interact. If a system has the “right” kind of causal structure, in other words, if its elements interact in the “right” way, it is conscious. Otherwise, it is not. We call such theories *causal structure theories*. For example, in Recurrent Processing Theory (RPT), Lamme proposed that recurrent processing is both necessary and sufficient for consciousness (Lamme, 2006). The first sweep of visual feedforward processing is unconscious. Consciousness kicks in when recurrent, top-down processing interacts with neurons activated during the initial feedforward sweep (Lamme, 2006). According to RPT, what matters is the causal structure because consciousness depends only on how neurons interact with each other: when there is recurrent processing there is consciousness, and there is no consciousness otherwise. Empirical support for RPT was proposed to be provided by neurophysiological experiments (Fig. 1c) in which recurrent processing enhanced neural activity in V1 when visual stimuli were consciously perceived. When the stimuli were not consciously perceived (during anaesthesia or when the stimuli were masked), there was no recurrent processing (Fahrenfort, Scholte, & Lamme, 2007).

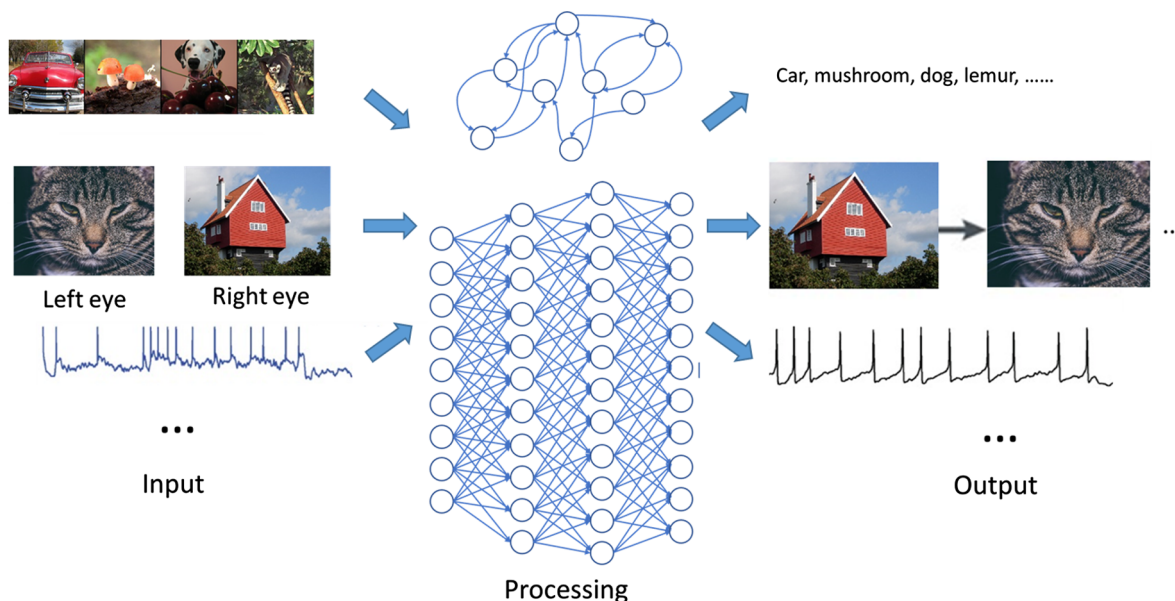
Information Integration Theory (IIT) is another example of a causal structure theory of consciousness. IIT proposes that an information integration measure called  $\phi$ , which is computed based on the causal structure of a system, quantifies consciousness (Oizumi, Albantakis, & Tononi, 2014). Consciousness is identified with  $\phi > 0$  systems: if elements of a system interact in the “right” way, the system has  $\phi > 0$  and is conscious. If  $\phi = 0$ , it is unconscious. For example,  $\phi$  is always greater than zero in recurrent systems (they are always conscious) and always equal to zero in feedforward systems (they are never conscious). Empirical support for IIT was asserted to be provided by studies showing that a practical proxy of  $\phi$  is low in coma, intermediate in minimally conscious states, and maximal during wakefulness (Casali et al., 2013; Tononi, Boly, Massimini, & Koch, 2016).

IIT and RPT were amongst the first theories of consciousness to make precise predictions about which systems are conscious. As such, they contributed greatly to the advancement of the science of consciousness. However, we will show that causal structure theories end up in an empirical impasse for principled reasons: they are either false or outside the realm of science.

## 2. The unfolding argument

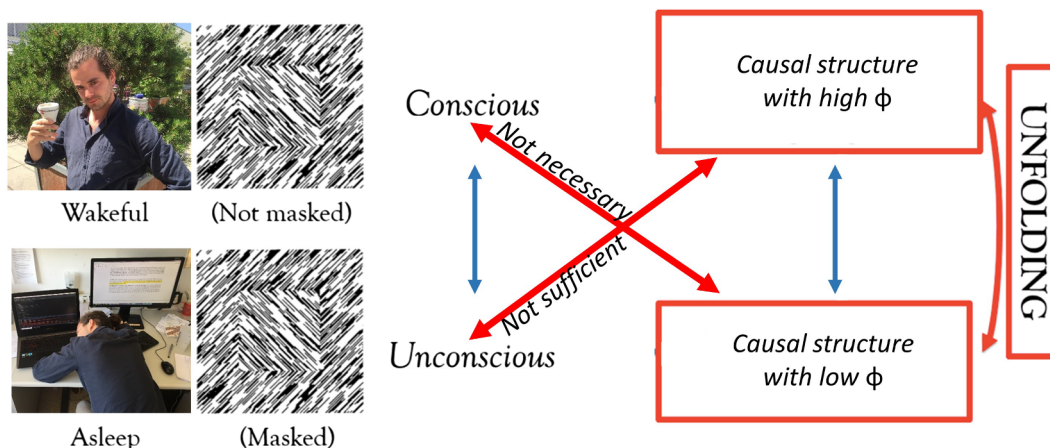
Recurrent neural networks are universal function approximators (Fig. 2; Schäfer & Zimmermann, 2006). That is, any input-output function can be approximated to any degree of accuracy. Vision is such an input-output function. For example, pictures of animals are presented as inputs on the retina, and the outputs are the elicited percepts of animals (or reports about these percepts). Likewise, the stimuli in a visual masking experiment are inputs, and the outputs may be button presses, verbal reports or any other measure shown to reliably correlate with subjective reports. Importantly, experiments which intervene directly on the brain, for example using implanted electrodes or Transcranial Magnetic Stimulation (TMS) are still input-output functions. The only difference is that part of the input is provided by means of electrodes or TMS rather than through the sensory organs.

Feedforward neural networks are also universal function approximators (Fig. 2; Hornik, Stinchcombe, & White, 1989). Hence, for



**Fig. 2.** Universal approximators & unfolding. Both recurrent networks (middle top) and multilayer feedforward networks (middle bottom) are universal function approximators (Hornik et al., 1989; Schäfer & Zimmermann, 2006). That is, they can be used to generate any desired input-output function to any degree of accuracy using a finite number of neurons. Therefore, for any recurrent network with a given input-output behaviour, there are corresponding feedforward networks with the same characteristics (although feedforward networks often need many more neurons than their recurrent counterparts). For example, recurrent networks performing image recognition, exhibiting binocular rivalry, and processing spike trains all have feedforward equivalents (see main text). Anything that can be done by recurrent networks can also be done in a feedforward manner.





**Fig. 3.** Double dissociation. Causal structure theories aim to explain empirical data about consciousness. For example, IIT proposes that  $\phi$  increases gradually as we go gradually from unconscious to conscious states, e.g., from sleep to drowsiness to wakefulness (vertical blue arrows). However, the unfolding argument shows that we can completely reverse the picture and, for example, implement conscious states with  $\phi = 0$  neural networks and unconscious states with high  $\phi$  neural networks. Hence,  $\phi$  is neither necessary (wakefulness can be implemented with  $\phi = 0$ ; downwards oblique red arrow) nor sufficient (sleep can be implemented with high  $\phi$ ; upwards oblique red arrow) to explain experimental results about consciousness. The same reasoning applies to RPT's figure-ground experiments (Fahrenfort et al., 2007). Therefore, causal structure is doubly dissociated from empirical data, i.e., it is neither necessary nor sufficient to account for experiments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a given input-output function we can find both feedforward and recurrent networks that realize the same function in different ways (LeCun, Bengio, & Hinton, 2015; Oizumi et al., 2014; Werbos, 1988). For instance, if there is a recurrent network that performs image recognition, there is an equivalent feedforward network that does it equally well. If there is a recurrent network that exhibits the characteristics of binocular rivalry, there is an equivalent feedforward network that does so too. If there is a recurrent network that takes a collection of spike trains as input and outputs another collection of spike trains, there is an equivalent feedforward network that does the same thing. Anything that can be done by recurrent networks can also be done in a feedforward manner (Fig. 2). We call this *unfolding*: any recurrent network can be unfolded into a feedforward network implementing the same function. In particular, any behavioural experiment can be seen as an input-output function, and can thus be implemented by both recurrent and feedforward networks.

Any input-output behaviour can be implemented not only by one particular feedforward network, but also by infinitely many equivalent feedforward networks and by infinitely many equivalent recurrent networks, because the universal approximator property does not depend on structural details such as the number of layers or on the precise connectivity. In fact, given an input-output function, we can find infinitely many networks, *each with a different*  $\phi$ , that all realize the same input-output function (see Appendices A and B, see also Chalmers, 2018). Moreover, the universal function approximator property is not restricted to neural networks but also holds true for Turing machines, cellular automata, cyclic tag systems, and more generally for any universal computing system (Turing, 1937; Wolfram, 2002). These facts are uncontroversial and widely accepted (including by proponents of IIT: see Oizumi et al., 2014).

### 2.1. Implication I: causal structure theories are doubly dissociated from empirical data

An implication of the unfolding argument is that causal structure theories are either false or outside the realm of science. In other words, causal structure theories are doubly dissociated from empirical data: they are neither necessary nor sufficient to explain empirical data (see Fig. 3). For example, according to IIT, the level of consciousness varies with  $\phi$ . A system is conscious if, and only if,  $\phi > 0$  (oblique red arrows in Fig. 3).<sup>1</sup> An experiment can be seen as an input-output function (Fig. 2). Hence, for any recurrent system with  $\phi > 0$  that reproduces the outcome of an experiment, there are feedforward systems with  $\phi = 0$  that also reproduce the outcome. According to IIT, one system has consciousness but the other does not. Conversely, for any feedforward system with  $\phi = 0$ , there are recurrent systems with  $\phi > 0$  that produce the same experimental results. In fact, we show in Appendix A how to implement any function with  $\phi = 0$  or with arbitrarily high  $\phi$ . That is to say, for each system that provides evidence for IIT, there are other possible systems that falsify it.

This argument generalizes to any causal structure theory, not just IIT. All causal structure theories are doubly dissociated from empirical results about consciousness. Hence, it makes no sense to provide experimental evidence for causal structure theories. For example, the figure-ground experiment mentioned earlier cannot support or be explained by RPT because there are feedforward

<sup>1</sup> In fact, IIT makes an even stronger claim:  $\phi$  varies monotonically with the degree of consciousness. The larger  $\phi$ , the stronger consciousness. Coma patients have low  $\phi$ , sleep comes with a higher  $\phi$ , and wakefulness with an even higher  $\phi$ .

networks that make the same subjective reports as humans when they consciously see the figure, but are unconscious according to RPT. Conversely, there are networks with recurrent activity that make the same subjective reports as humans when they do *not* consciously perceive the figure, but are conscious according to RPT. Likewise, the finding that awake humans have higher  $\phi$  than sleeping humans cannot be explained by or support IIT because there are feedforward networks with human wakefulness characteristics, and recurrent network with human sleep characteristics.

Our arguments are not only of an abstract mathematical nature. In real life, there are many examples where feedforward and recurrent networks realize the same complex functions. For example, deep reinforcement learning has been implemented with purely feedforward convolutional networks to achieve super-human performance in Atari video games (Mnih et al., 2013). Hausknecht and Stone (2015) replicated this superhuman performance using recurrent networks. The unfolding theorems tell us that this is not surprising because we can always find equivalent feedforward and recurrent networks. These systems are empirically identical (to a close approximation). One is conscious but the other is not, according to causal structure theories. Moreover, unfolding provides a *recipe* to build two small robot systems with *exactly* identical input-output functions but different causal structure (see Appendices A and B). Experiments on one robot support the theory; experiments on the other falsify it.

The unfolding argument shows that there are always systems that empirically falsify causal structure theories. Proponents of IIT try to avoid this problem by claiming that systems with  $\phi = 0$  are unconscious *despite being empirically indistinguishable from conscious systems* (Oizumi et al., 2014). We will show next that this claim makes IIT circular, and therefore unfalsifiable. In other words, causal structure theories are falsified by the unfolding argument, unless they decide to become unfalsifiable.

For example, proponents of IIT may still insist that the robot with  $\phi = 0$  is unconscious whereas the one with  $\phi > 0$  is conscious, owing to their differing causal structure. However, such a proposition quickly ends up in circularity because we have no criteria to settle the matter. In particular, we have no *empirical* criteria because experimental results about consciousness are all identical for the two robots. The only reason to believe that only the  $\phi > 0$  robot is conscious is to already believe in IIT, but this is circular. The situation is even worse: there are many causal structure theories, such as IIT and RPT. Which one is the “right” one? Even within IIT, the axioms do not uniquely determine  $\phi$  (Barrett & Mediano, 2019; Bayne, 2018), and different empirical measures of  $\phi$  yield very different results (Mediano, Seth, & Barrett, 2018). Which version of IIT is the “right” one? We can never decide because we have no criteria to test the theories and pit their predictions against each other. We are left with the conclusion that there are different types of “consciousness” (i.e.,  $\text{consciousness}_{\text{IIT\_version}_1}, \dots, \text{consciousness}_{\text{IIT\_version}_n}, \text{consciousness}_{\text{RPT}}$ , and so on), depending on which theory we favour. Insisting that the robot with  $\phi = 0$  is unconscious whereas the one with  $\phi > 0$  is conscious even though they are empirically identical leads IIT outside the realm of empirical science.

To summarize the unfolding argument, the conclusion follows from four premises.

**(P1):** In science we rely on physical measurements (based on subjective reports about consciousness).

**(P2):** For any recurrent system with a given input-output function, there exist feedforward systems with the same input-output function (and vice-versa).

**(P3):** Two systems that have identical input-output functions cannot be distinguished by any experiment that relies on a physical measurement (other than a measurement of brain activity itself or of other internal workings of the system).

**(P4):** We cannot use measures of brain activity as a-priori indicators of consciousness, because the brain basis of consciousness is what we are trying to understand in the first place.

**(C):** Therefore, EITHER causal structure theories are falsified (if they accept that unfolded, feedforward networks can be conscious), OR causal structure theories are outside the realm of scientific inquiry (if they maintain that unfolded feedforward networks are *not* conscious despite being empirically indistinguishable from functionally equivalent recurrent networks).

## 2.2. Examples

Imagine that one could surgically replace the brain’s native recurrent sound processing system with an equivalent feedforward implant. The implant takes the same collection of spike trains as inputs, and outputs the same collection of spike trains as the native brain areas. We know that such implants exist in principle because of the previously mentioned unfolding theorems. Even though the causal structure in the new implant is completely different, the rest of the brain does not notice any difference.<sup>2</sup> The brain can do its normal job. This means that all subjective reports by the person are identical before and after the surgery. The person will claim all the same things about sound as before the implant was placed, such as “I hear the drizzle of the rain, it is music to my ears”, or “I understand what you are saying”, etc. In particular, any experiment about which sounds are consciously perceived will yield exactly the same results as with the native brain area. Therefore, we end up with the dilemma mentioned earlier: either causal structure theories are wrong (if they accept that there is still auditory consciousness with the implant), or they are outside the realm of science (if they claim that consciousness is different with and without the implant even though there are no empirical differences).

We can push the example further to entire brains. Since anything that can be done with a recurrent network can also be done with a feedforward network, there could be «feedforward brains» that behave exactly like human brains. Such systems would have all the same functional characteristics as a normal human brain, but completely different causal structure. They behave exactly like a human in *all* respects, passing the Turing test seamlessly. However, according to causal structure theories, they are not conscious because

<sup>2</sup> This implies a certain level of functional isolation of the sound processing system, which may or may not be the case. However, our argument goes through regardless. Indeed, even if sound processing areas cannot be isolated, unfolding is applicable to the whole brain, as we illustrate below.

they do not have the “right” kind of causal structure.

Crucially, these systems respond to any empirical experiment *exactly like humans*. For example, they identically describe what it is like for them to see red, hear sounds, have memories, and so on. They respond to all scientific paradigms (such as masking, binocular rivalry, figure-ground segmentation, etc.) in exactly the same way. They exhibit the same wakefulness characteristics and the same sleep characteristics. In summary, no behavioural experiment can distinguish between human brains and feedforward brains in principle. Therefore, either causal structure theories are wrong or they are outside the realm of science.

### 2.3. Implication II: conscious content in IIT is doubly dissociated from experiments

In causal structure theories the *content* of consciousness is also doubly dissociated from empirical observations. For example, we can construct systems that behave as having experience X when, according to IIT, they are in fact experiencing Y (see [Appendix C](#)). For example, a system participating in a rivalry experiment may report that it is seeing the cat image when, according to IIT, it is experiencing the smell of ham. In principle, as shown in the appendix, it can experience *any* content of consciousness while reporting that it sees a cat. Of course, it could also experience seeing a cat, but this would just be a coincidence, showing a double dissociation.

There is a straightforward reason why causal structure theories are vulnerable to the kind of arguments presented here. All that is required for a system to be conscious is a particular causal structure. At the same time, any function can be implemented by many different systems with different causal structures. Hence, there can be no consistent link between causal structures and experimental results.

### 2.4. Network efficiency & evolutionary constraints

In practice, the brain has to cope with very strong space and energy constraints: processing must be efficient enough to be contained within a small skull, and energy consumption must be limited. Experiments have suggested that, indeed, sufficiently complex tasks can strongly constrain network properties, when the number of neurons is limited ([Khaligh-Razavi & Kriegeskorte, 2014](#); [Nayebi et al., 2018](#); [Yamins et al., 2014](#)). In general, feedforward networks require many more neurons to implement a function than equivalent recurrent networks with more efficient causal structure and are therefore impractical (but not always: for instance image recognition is more efficiently implemented in feedforward convolutional networks). In this regard, causal structure theories may turn out to be good markers for consciousness. For example, high  $\phi$  has obvious functional benefits, such as efficiently integrating information. We argue that awake brains have high  $\phi$  for this functional reason. Hence, causal structures may be good correlates for consciousness in humans not because they are identical with consciousness, but because they correlate well with neural information processing in general, which happens to covary with conscious state in humans as a contingent rule. This explains why causal structure theories may provide human consciousness-meters (see for example [Casali et al., 2013](#)). However, it is an entirely different thing to *identify* consciousness with causal structure. In short, brains are recurrent because brain processing must necessarily fit inside a skull, not because consciousness is identical with the brain’s causal structure.

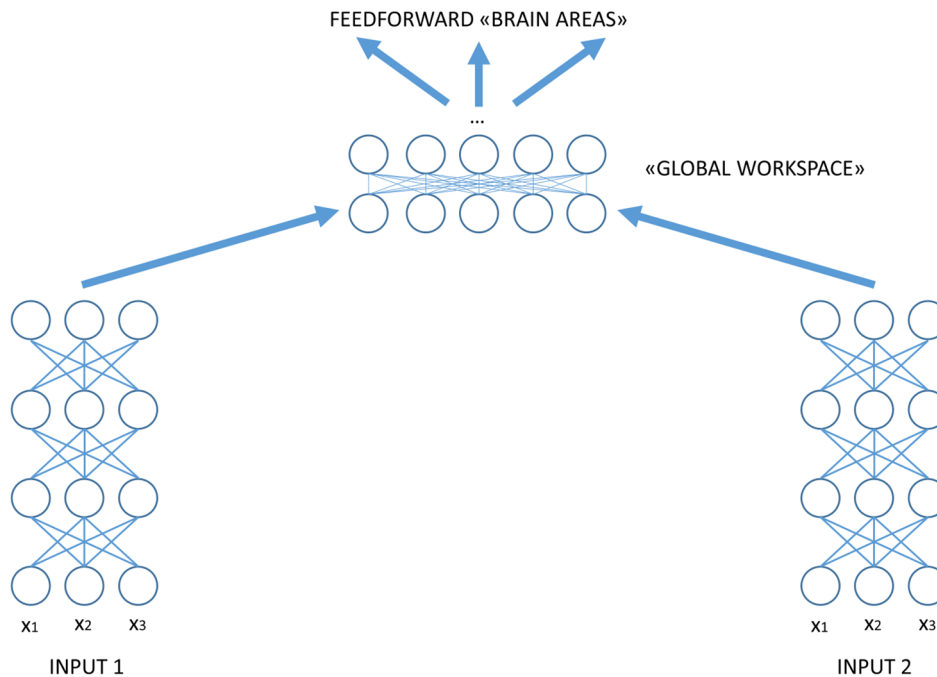
### 2.5. The unfolding argument vs. the zombie argument

The zombie argument is a well-known argument aiming to show that physicalist and functionalist theories cannot account for consciousness. Imagine unconscious “zombies” who are empirically indistinguishable from conscious “non-zombie” equivalents. For example, if a zombie inadvertently places its hand on a hot stove, it yells “ouch” and immediately retracts its hand, just as the non-zombie does. However, the zombie does not feel pain. The 3rd person, observable properties are all identical between zombies and non-zombies, but consciousness differs. Whether or not zombies are in fact possible is heavily debated (e.g., [Dennett, 1991](#)). However, if they are, it follows that consciousness cannot be explained in the standard, functionalist framework of science (this is the hard problem of consciousness; [Chalmers, 1996](#)). Indeed, if functionally identical systems (the unconscious zombie and its conscious non-zombie equivalent) can have different consciousness, then functional approaches cannot explain consciousness.

The unfolding argument is very different from the zombie argument for two reasons. First, the zombie argument aims to dismiss all physicalist accounts of consciousness, including functional ones. In contrast, the unfolding argument only targets causal structure theories of consciousness, and not physicalist or functionalist theories in general. In fact, the unfolding argument favours functionalist theories because (un)folded a network changes only its causal structure but not its function or physical nature. Other major theories of consciousness, such as Global Workspace Theory (GWT; [Baars, 1997](#); [Dehaene & Naccache, 2001](#)), Higher-Order Thought Theory (HOTT; [Lau & Rosenthal, 2011](#); [Rosenthal, 2004](#)) or Predictive Processing Theory (PPT; [Friston, 2013](#)) are not affected by the unfolding argument, as we show in the next subsection.

Second, one can choose to dismiss the zombie argument by claiming that zombies are in fact not possible (e.g., [Dennett, 1991](#)). In contrast, the existence of unfolded systems is a straightforward mathematical fact, and not a mere thought experiment. In fact, unfolding provides a recipe for creating empirically identical networks with different causal structures (for example, with arbitrarily high  $\phi$ ; see [Appendix A and B](#)). As mentioned, there even are *real-world* cases of feedforward and recurrent agents performing the same complex task (see [Section 2.1](#)). Hence, for example, the fact that we never have observed an unfolded cortex in practice (and probably never will) is not by itself a sufficient argument to call into question the unfolding argument. Furthermore, even though unfolded brains are impractical, we explicitly showed that the unfolding argument does not rely only on unfolded whole brains (see the previous example with the auditory system). This example can be scaled down again to the smallest part of the brain proposed to be relevant for consciousness by a given causal structure theory.





**Fig. 4.** A feedforward toy model of Global Workspace Theory. First, two feedforward networks process incoming sensory information, representing for example two different sensory modalities. Both project to the “global workspace”, which is simply another feedforward network maintaining activity through time by copying each layer to the next. At any time step (i.e., at any layer), the global workspace may be “queried” by one or several other processes, thus making the information globally available for other areas (broadcasting). Each of these other areas can be implemented simply by applying one feedforward network implementing the relevant function to the “global workspace”. Hence, this model fulfills the crucial functions proposed by GWT with a different causal structure, and the unfolding argument does not apply. There are many other ways to implement GWT in a feedforward fashion too. These networks can be “folded” back into recurrent networks and retain the same crucial functions. Hence, GWT can be implemented equivalently in recurrent and feedforward networks and does not face the unfolding argument.

## 2.6. Non-causal structure theories of consciousness are not subject to the unfolding argument

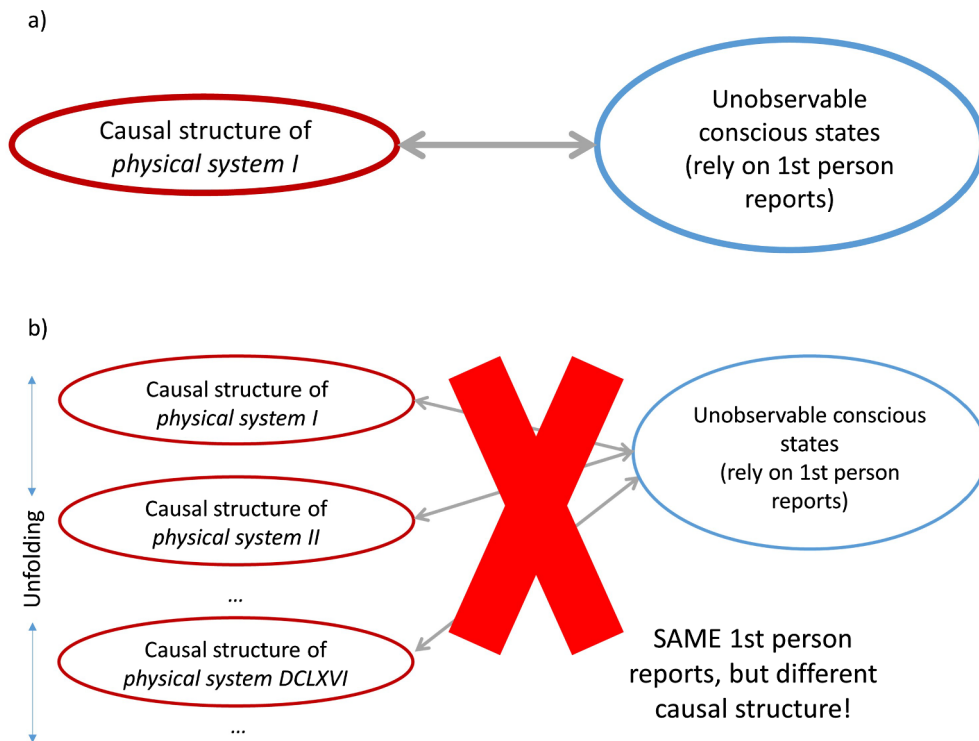
Global Workspace Theory (GWT; Baars, 1997; Dehaene & Naccache, 2001), Higher-Order Thought Theory (HOTT; Lau & Rosenthal, 2011; Rosenthal, 2004) or Predictive Processing Theory (PPT; Friston, 2013) are examples of functionalist theories of consciousness: they focus on *functions* proposed to be crucial for consciousness. The unfolding argument does not apply to these theories because they propose that systems are conscious insofar as they implement the right kind of function – independently of the causal structure. Of course, these theories are usually couched in terms of recurrent or top-down processing, or other seemingly causal-structure terminology, but they can be formulated in other kinds of networks too. The unfolding argument only applies to theories in which recurrence *per se* (or another proposed causal structure) is necessary and sufficient for consciousness.

For example, the typical description of GWT is that consciousness occurs when cortical areas, which code for certain contents of consciousness (e.g., sensory areas), “broadcast” their information in a global neuronal workspace that consists of highly *recurrent* fronto-parietal areas, thus making these contents globally available for widespread use by other areas. The crucial functions here are (a) the creation of contents in sensory areas and (b) making these contents globally available for widespread use by other areas (i.e., broadcasting). GWT is usually explained with recurrent networks. Still, equivalent feedforward networks can maintain the same broadcasting function (see Fig. 4 for a toy model).<sup>3</sup> Similar toy models are easily produced for HOTT, PPT and all other functionalist theories. What matters is the function, e.g., broadcasting, but not the (neural) implementation. In summary, functionalist theories differ importantly from causal structure theories in that they propose functions as crucial for consciousness, independently of their implementations. For example, an unfolded global workspace network retains the crucial function of broadcasting. Hence, by their very nature, functionalist theories are not subject to the unfolding argument.

## 2.7. Unfolding & the correlation approach

For many researchers, consciousness is non-physical and cannot be studied using input-output functions, and therefore might be impossible to explain with standard neuroscience (Chalmers, 1996). In this case, it is impossible to study consciousness directly. However, consciousness may still be linked to neural states by bridging principles based on correlations (Chalmers, 2004; Varela,

<sup>3</sup> Alternatively, if GWT claims that recurrence *per se* is essential for consciousness, it becomes a causal structure theory and the argument applies.



**Fig. 5.** (a) The correlation approach proposes that we can find correlations between physical properties (such as the brain’s causal structure, left) and first person reports about conscious states (right). If this is true, we can have a theory of consciousness even if conscious states are taken as unobservables. (b) However, the unfolding argument shows that there are infinitely many equivalent systems leading to the same first person reports but with different causal structures. Therefore, the correlation approach cannot be used to link causal structure theories with empirical data.

1996). For example, we may find correlations between human reports about their conscious experience (first-person data: I am experiencing a face) and observable properties of the brain (third person data: neural activity in the fusiform face area). To quote Chalmers (2004): «In the case of consciousness, we can expect systematic *bridging principles* that underlie and explain the covariation between third-person data and first-person data.»

This approach cannot hold for causal structure theories because of the unfolding argument (Fig. 5). As mentioned, there are infinitely many equivalent systems that produce exactly the *same* first person reports as humans, but with completely *different* causal structures. Therefore, linking conscious properties with the brain’s causal structures by relying on first person reports cannot succeed (Fig. 5). The correlation approach cannot work with causal structure theories, although it may (or may not) succeed for other theories of consciousness.

### 3. Concluding remarks

To be considered scientific, IIT and other causal structure theories require empirical support. However, the unfolding argument shows that they are either false or outside the realm of science. For the same reason, different causal structure theories cannot be compared with each other. For example, different mathematical formulations of IIT’s axioms lead to different predictions about which systems are conscious, but we cannot compare them because the predictions are doubly dissociated from empirical data. Proponents of IIT have previously acknowledged that feedforward and recurrent networks can be functionally equivalent but have different consciousness, according to IIT (Oizumi et al., 2014). In other words, they share the same uncontroversial starting point as we do. However, conclusions differ strongly. Proponents of IIT suggest that this should prompt us to focus on the subjectivity of consciousness. In contrast, we conclude that adopting a causal structure theory precludes *any* experimental approach to consciousness. Indeed, we have shown that *all* possible experimental results, including the ones focussing on subjectivity, do not depend on causal structure.

The unfolding argument rules out a class of explanations of consciousness wherein consciousness supervenes on causal structures. This should prompt us to turn our attention elsewhere in trying to understand consciousness. In this respect, the unfolding argument suggests that consciousness must be explained on a more abstract level than that of neural wiring. Indeed, any proposed framework based on neural connections suffers from the unfolding argument: any network can be replaced by equivalent feedforward networks with different connections that lead to identical empirical observations about consciousness. Only theories that abstract away implementation details and focus on explaining which kinds of functions are important for consciousness can avoid these challenges. To remain within the realm of science, consciousness must be described in terms of what it does, and not how it does it.

## Acknowledgements

We thank Matthias Michel and Scott Aaronson for helpful discussions. This work was supported by the SNF grant “Basics of visual processing: from elements to figures” (176153).

## Author contributions

All authors contributed to developing the unfolding argument and its implications. AD & MHH wrote the manuscript. AS & KH provided feedback and improvements.

## Appendix A. All functions can be implemented with an arbitrarily high level of consciousness ( $\phi$ )

Aaronson showed (Aaronson, 2014) and Tononi agreed (Tononi, 2014) that certain kinds of systems can have arbitrarily high  $\phi$ . We call these systems  $\phi$ -increasing devices. Examples include expander graphs, XOR grids, Vandermonde matrices, large random networks, and others. We will show that these  $\phi$ -increasing devices can be inserted into feedforward networks without altering their function.

Let us consider a feedforward network implementing the function  $x \rightarrow F(x)$ . Here,  $x$  represents the input to the system, which is mapped to the output  $F(x)$ . Being feedforward, this network has  $\phi = 0$ . We can change the network's  $\phi$  to arbitrarily high values by inserting a  $\phi$ -increasing device in the network, and subsequently canceling its functional effect (Fig. A1a). This is proven by noticing that a  $\phi$ -increasing device can be described as applying a function  $g$  to its input. Since feedforward networks are universal function approximators, there are feedforward networks implementing the inverse,  $g^{-1}$ , thereby cancelling the functional effect of the  $\phi$ -increasing devices. The function  $g^{-1}$  exists at least for certain  $\phi$ -increasing devices. For example, Aaronson showed that a system recursively applying Vandermonde matrices is a  $\phi$ -increasing device (Aaronson, 2014), and Vandermonde matrices are invertible. Adding a  $\phi$ -increasing device increases the  $\phi$  of the entire network. In fact, the network has exactly the  $\phi$  of the  $\phi$ -increasing device because the feedforward parts do not add to  $\phi$ . Applying the inverse function  $g^{-1}$  cancels  $g$  but *does not change*  $\phi$  (Fig. A1a).<sup>4</sup> Hence, the input-output functions of the original feedforward network and the modified network with the  $\phi$ -increasing device are the same but the modified network can have an arbitrarily high value of  $\phi$  depending on the device. These systems with arbitrarily high  $\phi$  are indistinguishable by experiments about consciousness.

## Appendix B. Other implementations with arbitrarily high $\phi$

It is known from the theory of computation that any given function can be implemented in infinitely many different ways. In particular, there are many alternative systems that implement a function with arbitrarily high  $\phi$ .

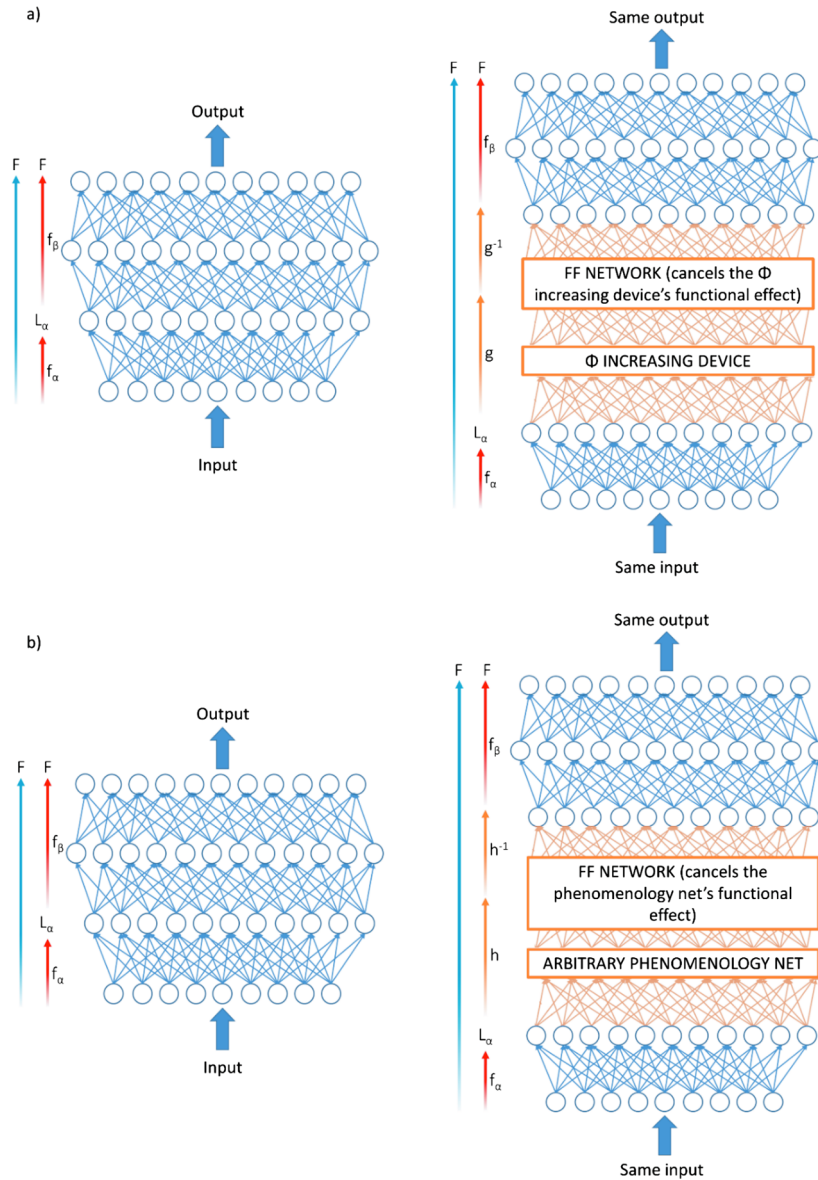
For example, we can use echo state networks (also known as liquid state machines, or reservoir computing; Jaeger, 2001; Maass, Natschläger, & Markram, 2002). Inputs drive the dynamics of a recurrent population of neurons (called the “reservoir”). Outputs are then decoded from the reservoir dynamics. Feeding inputs to and decoding outputs from the reservoir can be done in a feedforward fashion and therefore do not contribute to  $\phi$ . Echo state networks are also universal function approximators, provided the reservoir is large enough and has sufficiently rich dynamics (Maass et al., 2002). Often, the reservoir is a sparsely, randomly connected recurrent network, but it has been shown that the randomness is not crucial, all that is needed is the possibility of sufficiently rich dynamics (Rodan & Tino, 2011). Hence, in principle, it is possible to train an echo state network to perform any function using a  $\phi$ -increasing device as a reservoir. Since the reservoir has arbitrarily high  $\phi$  and can be used to learn any function, any function can be implemented with arbitrarily high  $\phi$ .

## Appendix C. All functions can be implemented with arbitrary content of consciousness (shapes in qualia space)

In IIT, the part of the system (called a complex) with maximum  $\phi$  determines the content of consciousness, also called a quale. A complex, non-computable procedure maps the connectivity and state of this maximal  $\phi$  complex onto the corresponding quale (Balduzzi & Tononi, 2009). For example, a certain brain activity may have a maximal  $\phi$  complex corresponding to the quale of smelling ham. The quale of a composite system, consisting of several feedforward networks and one recurrent network (such as depicted in Fig. A1b), is equal to the quale of the recurrent network alone.

Let us consider a feedforward network, which performs a binocular rivalry task (Fig. 1). Being feedforward, the network has no consciousness and hence no content of consciousness. We now use the same trick as before but, instead of inserting a  $\phi$ -increasing device, we insert a recurrent network whose states and connectivity correspond to a certain quale, for example, the quale of smelling ham. Next, we cancel the effect of the recurrent network with a feedforward network. The recurrent network is the maximal  $\phi$  complex because the rest of the network is feedforward. Hence, the network experiences the smell of ham consciously but reports performing a rivalry experiment (for example it reports seeing the cat image from Fig. 1 consciously). We can replace the quale of

<sup>4</sup> The specifics of how to compute  $\phi$  are beyond the scope of this appendix. For present purposes, it is sufficient to point out that the  $\phi$  value of a composite system consisting of several feedforward networks and one recurrent network (such as depicted in supplementary Fig. A1) is equal to the  $\phi$  value of the recurrent network alone. Feedforward networks do not contribute to  $\phi$  (Oizumi et al., 2014).



**Fig. A1.** (a) Networks with identical function and arbitrary  $\phi$ . *Left:* A feedforward network implementing a function  $F$  with  $\phi = 0$ . *Right:* A network implementing the same function  $F$  with arbitrarily high  $\phi$ . In the “original” feedforward network (left), the function  $F$  can be decomposed into two successive functions: a first function  $f_\alpha$  from the network input to layer  $L_\alpha$ , and a second function  $f_\beta$  from  $L_\alpha$  to the output layer. We have  $f_\beta \circ f_\alpha = F$ , with  $\phi = 0$ . Now, we increase  $\phi$  by inserting a  $\phi$  increasing device after layer  $L_\alpha$  and we subsequently counteract its functional effect using a feedforward network (orange neurons). In the modified network (right), there is the same function  $f_\alpha$  from the network inputs to activities of layer  $L_\alpha$  (bottom blue part). Then, processing continues in the  $\phi$  increasing device, which implements a function  $g$ .  $g$  is cancelled by a subsequent feedforward network, which implements the function  $g^{-1}$  ( $g^{-1}$  exists on the image of  $g$  if  $g$  is injective). These steps are represented in orange. Lastly, the rest of the network is unchanged and implements the function  $f_\beta$  from the output of this entire process to the networks output (upper blue part). We have  $f_\beta \circ g^{-1} \circ g \circ f_\alpha = F$ , with an arbitrarily high  $\phi > 0$ . The  $\phi$  of the network is entirely determined by the  $\phi$  increasing device because all other parts of the network are feedforward. (b) Networks with identical function and arbitrary phenomenology. *Left:* The same feedforward network as in a), implementing the function  $F$  with  $\phi = 0$  (so there is no conscious content). We have  $f_\beta \circ f_\alpha = F$ , with no conscious content. *Right:* Because IIT provides a principled way to determine which connectivities and states produce which qualia, we can slightly modify the example above to give arbitrary qualia to networks with the same function. We play the same trick as before but, instead of inserting a  $\phi$ -increasing device after layer  $L_\alpha$ , we insert a network proposed by IIT to produce certain qualia (for example, the quale of smelling ham). This network applies the function  $h$  to the outputs of layer  $L_\alpha$ . Just as before, the function can be inverted by a feedforward network implementing  $h^{-1}$  ( $h^{-1}$  exists if  $h$  is injective). Overall, we now have  $f_\beta \circ h^{-1} \circ h \circ f_\alpha = F$ , with the quale of smelling ham (or any other quale). Using a feedforward network to cancel the functional effect of inserting our quale-of-smelling-ham network does not simultaneously abolish the quale of smelling ham because, being feedforward, it is not part of the maximal  $\phi$  complex.

smelling ham by any other quale, for example, by the quale of seeing a cat. However, this would be just a coincidence, showing that conscious percepts and input-output functions are doubly dissociated.

As mentioned, any function can be implemented by many different systems, with different causal structures. Hence, there are many examples of systems performing the exact same function with arbitrary qualia.

## References

- Aaronson, S. (2014). Why i am not an integrated information theorist (or, the unconscious expander). Retrieved August 23, 2018, from <<https://www.scottaaronson.com/blog/?p=1799>>.
- Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4), 292–309.
- Balduzzi, D., & Tononi, G. (2009). Qualia: The geometry of integrated information. *PLoS Computational Biology*, 5(8), e1000462.
- Barrett, A. B., & Mediano, P. A. (2019). The phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, 26(1–2), 11–20.
- Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 2018(1), niy007.
- Casali, A. G., Gossesies, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., ... Tononi, G. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198) 198ra105–198ra105.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Chalmers, D. J. (2004). *How can we construct a science of consciousness?* MIT Press.
- Chalmers, D. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9–10), 6–61.
- Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1), 1–37.
- Dennett, D. C. (1991). *Consciousness explained*. New York: Little Brown & Co.
- Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, 19(9), 1488–1497.
- Friston, K. (2013). Consciousness and hierarchical inference. *Neuropsychanalysis*, 15(1), 38–42.
- Hausknecht, M., & Stone, P. (2015). Deep recurrent q-learning for partially observable mdps. *CoRR, Abs/1507.06527*, 7(1).
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.
- Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34), 13.
- Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11), e1003915.
- Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494–501.
- Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8), 365–373.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>.
- Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11), 2531–2560.
- Mediano, P.A., Seth, A.K., & Barrett, A.B. (2018). Measuring integrated information: Comparison of candidate measures in theory and simulation. ArXiv Preprint ArXiv:1806.09373.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. ArXiv Preprint ArXiv:1312.5602.
- Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... Yamins, D.L. (2018). Task-driven convolutional recurrent models of the visual system. ArXiv Preprint ArXiv:1807.00053.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, 10(5), e1003588.
- Rodan, A., & Tino, P. (2011). Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1), 131–144.
- Rosenthal, D. M. (2004). Varieties of higher-order theory. *Advances in Consciousness Research*, 56, 17–44.
- Schäfer, A. M., & Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. *Artificial Neural Networks– ICANN 2006*, 632–640. [https://doi.org/10.1007/11840817\\_66](https://doi.org/10.1007/11840817_66).
- Tononi, G. (2014). Why Scott should stare at a blank wall and reconsider (or, the conscious grid). Retrieved from <<http://www.scottaaronson.com/blog/?p=1823>>.
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450.
- Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2–42(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>.
- Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3(4), 330–349.
- Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, 100(10), 6275–6280.
- Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4), 339–356. [https://doi.org/10.1016/0893-6080\(88\)90007-X](https://doi.org/10.1016/0893-6080(88)90007-X).
- Wolfram, S. (2002). *A new kind of science*, Vol. 5. Wolfram media Champaign.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.