1-12-2022

First-Person Experience Cannot Rescue Causal Structure Theories from the Unfolding Argument

Michael H. Herzog
EPFL – École polytechnique fédérale de Lausanne

Aaron Schurger
Chapman University, schurger@chapman.edu

Adrian Doerig
Donders Institute for Brain, Cognition & Behaviour

Recommended Citation

Herzog, M. H., Schurger, A., & Doerig, A. (2022). First-person experience cannot rescue causal structure theories from the unfolding argument. *Consciousness and Cognition, 98*, 103261. https://doi.org/10.1016/j.concog.2021.103261

First-Person Experience Cannot Rescue Causal Structure Theories from the Unfolding Argument

Consciousness and Cognition 98 (2022) 103261

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

First-person experience cannot rescue causal structure theories from the unfolding argument

Michael H. Herzog [a,*], Aaron Schurger [b,c,d,e], Adrien Doerig [f]

[a] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale De Lausanne (EPFL), Lausanne, Switzerland
[b] Department of Psychology, Crean College of Health and Behavioral Sciences, Chapman University, Orange, CA, USA
[c] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Irvine, CA, USA
[d] INSERM, Cognitive Neuroimaging Unit, Gif sur Yvette 91191, France
[e] Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin, center, Gif sur Yvette 91191, France
[f] Donders Institute for Brain, Cognition & Behaviour, Nijmegen, Netherlands

A B S T R A C T

We recently put forward an argument, the Unfolding Argument (UA), that integrated information theory (IIT) and other causal structure theories are either already falsified or unfalsifiable, which provoked significant criticism. It seems that we and the critics agree that the main question in this debate is whether first-person experience, independent of third-person data, is a sufficient foundation for theories of consciousness. Here, we argue that pure first-person experience cannot be a scientific foundation for IIT because science relies on taking measurements, and pure first-person experience is not measurable except through reports, brain activity, and the relationship between them. We also argue that pure first-person experience cannot be taken as ground truth because science is about backing up theories with data, not about asserting that we have ground truth independent of data. Lastly, we explain why no experiment based on third-person data can test IIT as a theory of consciousness. IIT may be a good theory of something, but not of consciousness. We conclude by exposing a deeper reason for the above conclusions: IIT's consciousness is by construction fully dissociated from any measurable thing and, for this reason, IIT implies that both the level and content of consciousness are epiphenomenal, with no causal power. IIT and other causal structure theories end up in a form of dissociative epiphenomenalism, in which we cannot even trust reports about first-person experiences. But reports about first-person experiences are taken as ground truth and the foundation for IIT's axioms. Therefore, accepting IIT leads to rejecting its own axioms. We also respond to several other criticisms against the UA.

1. Introduction

Integrated Information Theory (IIT; Oizumi et al., 2014) and other Causal Structure Theories (CSTs), such as the recurrent processing theory (Lamme, 2006), propose that causal structure, i.e., how parts of a system interact rather than what the system does, is constitutive of consciousness. Here and previously, we argue that this proposal cannot hold true for principled reasons. In this contribution, we focus on IIT because it is the most widely known and elaborated CST. The very same arguments apply to any CST.

IIT starts from five "phenomenological axioms", proposed to provide ground truth about the essential properties of conscious experience, and asks which physical systems fulfill these essential properties. By mathematizing these axioms, IIT proposes that a certain measure of information integration based on causal structure, called Φ, indexes the quantity or magnitude of consciousness (Oizumi et al., 2014). Purely feedforward systems have $\Phi = 0$ and are hence never conscious. Recurrent systems have $\Phi > 0$ and are hence always conscious. The content of consciousness is determined by and can be computed from the causal structure (called "the qualia space", Balduzzi & Tononi, 2009; or, more recently, "the shape of the cause-effect structure", Haun & Tononi, 2019). IIT

; UA, Unfolding Argument; IIT, Integrated Information Theory; CST, Causal Structure Theory; i/o, input/output.

* Corresponding author.

https://doi.org/10.1016/j.concog.2021.103261
Received 3 July 2021; Received in revised form 29 October 2021; Accepted 7 December 2021
Available online 12 January 2022

purports to be empirically grounded.

In the empirical sciences, input–output (i/o) functions are behind all experiments. Physicists perturb a system (input) and measure how it reacts (output). Psychologists measure how people react (output) in different situations (input). Even high-level sciences such as political sciences strive to understand how political choices (inputs) affect outcomes (outputs).

Experiments on consciousness are no different (Dehaene et al., 2017; Doerig, Schurger, et al., 2020; Francken et al., 2021). Researchers study how different stimuli (inputs) lead to different reports about subjective experiences (outputs). *Any possible experiment about consciousness can be seen as measuring i/o functions.* In humans the inputs can be visual stimuli, the outputs can be verbal reports or button presses, and the i/o functions map the stimuli to their outputs. For example, in a binocular rivalry experiment, the inputs are the two images presented separately to the two eyes and the outputs are button presses about which of the two stimuli is perceived. Even cases where the role of i/o functions is not obvious are based on i/o functions. For example, we know that locked-in or minimally conscious patients are conscious because researchers were able to communicate with them using neuroimaging or other methods to gather the subjects' reports (Casali et al., 2013; Demertzi et al., 2019). In so-called no-report paradigms we solicit responses (outputs) before or after the no-report experiment, in order to infer what the participant was likely experiencing (or not experiencing) during the no-report experiment. For example, we may infer the conscious percept based on a reflex, like the optokinetic nystagmus, that has previously been validated by showing that it correlates with what the subjects report seeing (Tsuchiya et al., 2015). Importantly, this does *not* mean that i/o functions are "all there is" to consciousness. However, one cannot do scientific research entirely without i/o functions.

In a previous contribution, we proposed the unfolding argument (Doerig et al., 2019), which argues that IIT and other CSTs are either false or unfalsifiable under standard scientific practice, i.e., based on i/o function, because their consciousness is completely dissociated from experimental results about consciousness as a matter of logic. Well-known mathematical theorems prove that, for any feedforward neural network ($\Phi = 0$), there are recurrent networks ($\Phi > 0$) that have identical i/o functions and vice-versa (Hornik et al., 1989; LeCun et al., 2015; Oizumi et al., 2014; Schäfer & Zimmermann, 2006; Werbos, 1988). Hence, according to IIT, there can be two systems with identical i/o functions, i.e., all their observable behaviors are identical, but one is always conscious whereas the other never is.

To illustrate, we can build two robots with identical i/o functions but different Φ (Doerig et al., 2019; see Hanson & Walker, 2020 for a detailed real-life example). One robot is feedforward ($\Phi = 0$) and is never conscious according to IIT, and the other uses recurrent connections ($\Phi > 0$) and is always conscious. Therefore, IIT proposes that two systems can have different consciousnesses despite being identical for *all* i/o functions. Conversely, two robots can have the same Φ, but one shows all i/o signs of consciousness (for example responses in psychophysical studies routinely used in consciousness research) and the other robot's brain is just an inactive grid of XOR gates (Aaronson, 2014b). Hence, there is a full double dissociation between consciousness and i/o functions under IIT.

We summarized the Unfolding Argument (UA) as follows:

(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).

In other words, the results of experiments about consciousness *provably* do not depend on causal structure. When we observe an experimental result about consciousness, we know that causal structure is neither necessary nor sufficient to produce it.

Not surprisingly, our contribution provoked critical responses (Albantakis, 2020; Kent & Wittmann, 2021; Kleiner, 2020; Kleiner & Hoel, 2021; Mallatt, 2021; Negro, 2020; Tsuchiya et al., 2020; Usher, 2021) but also support (Ganesh, 2020; Hanson, 2021; Hanson & Walker, 2019, 2020; Hanson & Walker, 2021). For example, Hanson and Walker proved similar results using a different formalism (the Krone-Rhodes theorem; Hanson & Walker, 2019) and showed that feedforward and recurrent systems can be equivalent on a deeper computational level than just i/o functions (Hanson & Walker, 2020). They provided blueprints to build real-world systems that behave identically but have different consciousness according to IIT (Hanson & Walker, 2020).

Here, we address the main criticisms of the UA. We use IIT as an example because it has been central in several criticisms, but the arguments generalize to CSTs in general. In the Arguments section, we respond to the central criticism that first-person experience can be used to avoid the UA. First, we will discriminate between two senses of "first person experience": the notion of introspection commonly used in consciousness science, versus the kind of "*pure* first-person experience" that is needed to avoid the UA. We will show that pure first-person experiences are not scientific data and hence cannot be used as a foundation for IIT independently of third-person data. Second, we will argue that using pure first-person experience as the basis for ground truth is untenable. Third, we show that the empirical evidence proposed to support IIT does not in fact support IIT *as a theory of consciousness*. IIT may be a good theory of something, but not necessarily a theory of consciousness. Fourth, we show that IIT implies a form of dissociative epiphenomenalism, where both the level and the content of consciousness are fully detached from any behavior, including the kind of behavior upon which IIT's first-person axioms are built – thus invalidating first-person experience as a basis of IIT.

Before moving on to the main arguments, we sketch five additional criticisms, which we discuss in depth in the Detailed Replies section. In addition, we clarify some confusion caused by our use of the notion of falsification.

1.1. Falsification

It seems that we sowed some confusion by using the terms "falsified" and "falsification" (Kleiner & Hoel, 2021; Negro, 2020; Tsuchiya et al., 2020). For instance, the robot example given earlier is not exactly a *falsification* in the sense of a theory that makes a prediction, which then turns out to be wrong after an experiment was carried out. In fact, what we have shown logically is that IIT is either *false,* rather than *falsified,* or circular, if consciousness science is limited to standard scientific i/o observations. As mentioned, it is false because there is a logically provable double dissociation between IIT's consciousness and i/o observations. If one still proposes to trust IIT in the absence of i/o evidence, it is circular. It seems that there is agreement with this reasoning (for a formal proof, see Kleiner, 2020). As mentioned, to avoid this problem, most counter-arguments are based on the idea that there is more to consciousness science than i/o observations, namely causal structure.

1.2. Can we trust (un)folded systems?

Albantakis (2020) proposed that we should not experiment on and reason about systems whose conscious status is unclear. For example, the above robot experiment may not be well suited to test consciousness since we do not know whether robots have consciousness. However, this argument seems to rely on our misuse of the term "falsification". As just mentioned, our claim is not that the unfolded robot counts as a surprising experiment that suddenly falsifies IIT. Rather, the UA shows the undeniable, logical implications of IIT: consciousness is fully dissociated from third-person data. The robot example is just an illustration of this fact, not a piece of empirical evidence needed to back up the UA. We spell out this argument in Detailed Replies section DR1, along with other responses to this "trust" criticism.

1.3. Behaviorism

Tsuchiya et al. (2020) argued that, by claiming that consciousness science must rely on i/o observations, we are "advocating for a new mode of methodological behaviorism". This statement could not be farther from the truth. A behaviorist would claim that internal states are useless to understand the mind, if they would use the word "mind" at all (Graham, 2000; Watson & McDougall, 1929). In contrast, we take subjective states seriously and assume that we can learn about them through i/o observations. We cannot measure consciousness directly, but we can measure subjective reports (verbal or otherwise) and link them to brain activity. That is not behaviorism, it is just science, as conducted by the majority of researchers in the field. Further rebuttals of Tsuchiya et al.'s criticisms are offered in the Detailed Replies section DR2.

1.4. Blockheads and consciousness

By appealing to the well-known blockhead thought experiment (Block, 1981), Negro (2020) argued that i/o experiments might not be all there is to consciousness science (see also Usher, 2021). The blockhead thought experiment purports to show that two systems can have identical i/o functions but differ in intelligence. For example, a lookup table (called "the blockhead") may play chess at a grand master level without real intelligence, and Negro applies the same line of thought to consciousness. However, we will argue in section D3 that the cases of consciousness and intelligence are different, because we have good independent reasons to believe that the blockhead is not intelligent. However, we do not have good independent reasons to believe that a system that reproduces all of human i/o functions is unconscious.

1.5. Do all theories of consciousness suffer from the UA?

Another form of argument, the "substitution argument", proposes that the UA leads to the rejection of (almost) all theories of consciousness, and that we may therefore need to reject at least one of its premises (Kleiner, 2020; Kleiner & Hoel, 2021). Although intriguing, we will argue that these arguments apply only to causal structure theories, as discussed in the Detailed Replies section DR4. In addition, Ganesh (2020) provides a strong counter-argument to the substitution argument, proving that a broad class of functionalist theories is immune to it.

1.6. Recurrent networks cannot be unfolded

Usher (2021) argued that the mathematical theorems used in the UA in fact do not show that recurrent networks can be unfolded into functionally equivalent feedforward networks, attacking our premise 2. There are two main steps to his argument. First, he acknowledges that unfolding networks through time can lead to functionally equivalent feedforward networks. But he dismisses unfolded networks because they are too large. Second, he claims that the functional equivalence of feedforward and recurrent networks does not extend to dynamical settings, where temporal aspects are involved. However, in the Detailed Replies section DR5, we will argue that there is no reason to discard unfolded networks and that feedforward networks can also cope with dynamical settings. We will also explain why the specific feedforward and recurrent networks discussed by Usher are not relevant for the UA, since they are

not in fact functionally equivalent.

2. Main Arguments: Can first-person experience rescue IIT?

As mentioned, all sides of the debate (except Usher) seem to agree that the UA holds if consciousness science is strictly based on third-person data gathered by i/o experiments (Albantakis, 2020; Doerig et al., 2019; Hanson & Walker, 2020; Kleiner, 2020; Kleiner & Hoel, 2021; Tsuchiya et al., 2020). *Therefore, the majority of counterarguments to the UA propose that there is more to consciousness research than just i/o experiments: in contrast to other sciences, first-person experiences are needed* (Albantakis, 2020; Kleiner, 2020; Kleiner & Hoel, 2021; Negro, 2020; Tsuchiya et al., 2020; see also Chalmers, 1996; Goff, 2019). Indeed, if this were true, the UA would not undermine causal structure theories, because even though third-person experiments about consciousness are dissociated from causal structure, purely first-person data may not be.

For this reason, the discussion has become focused around premise P3 of the UA: "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)". In particular, the bracket of P3, "(other than a measurement of brain activity itself or of other internal workings of the system)", led to strong disagreement. As Kleiner (2020) correctly points out, when taking the brain's causal structure into account one can easily distinguish whether a system is unfolded or not, and so the UA does not hold. Hence, the real controversy focusses on whether causal structure can make the difference for consciousness, without impacting i/o observations.

We argued that determining consciousness based only on causal structure leads to circularity and eventually unfalsifiability. Tsuchiya et al. (2020), Negro (2020), Albantakis (2020) and Kleiner (2020) argue that no such independent criterion is needed because the ground truths in consciousness science are one's own conscious experiences. Negro agrees that determining consciousness based only on Φ in the robot example would be circular, but he argues that the consciousness of a system is not directly defined by Φ, but rather derived from IIT's axioms, which carry undeniable ground truth in the form of first-person experience. For example, first-person experience delivers the ground truth that consciousness is integrated, which is captured by IIT's integration axiom. Oizumi et al. (2014) explicitly include Cartesian first-person truths as the basis for IIT. Thus, the argument goes, the axioms, and eventually Φ, are well-grounded in first-person truths and are therefore not circular. To put it differently, purely first-person experience can be used instead of third-person data to justify IIT.

2.1. Pure first-person experiences are not scientific data

To explore this possibility, let us start by clarifying some terminology. "Third-person data" denotes any data that we can observe and/or measure empirically, including verbal reports, button presses, imaging data, skin conductance, reaction times, or any other measure that can be carried out (Chalmers, 1996; Cohen & Dennett, 2011; Dehaene et al., 2017; Doerig, Schurger, et al., 2020). Third-person data can be shared with and observed by others. It is through confronting theories with publicly observable, sharable data that scientists can reach collective agreement. Non-sharable data is scientifically inert. Third-person data can be extremely rich and detailed, and, as mentioned, are at the basis of all sciences. When it comes to consciousness, all measurements must ultimately be grounded in subjective reports (button presses, verbal reports etc.). For example, the reason why it is legitimate to quantify the level of consciousness based on neuroimaging (Casali et al., 2013; Demertzi et al., 2019) is that we have first established that our neuroimaging signatures are correlated with the level of consciousness *by appealing to subjective reports.*

As mentioned, critics all claim that IIT can be grounded in "first-person experience" and thereby evade the UA. Here, we need to distinguish two notions of "first person experience". First, there is the notion of introspection commonly used in consciousness science. Of course, all psychophysical experiments are based on participants' reporting what they introspect about their first-person experiences. Did they experience the stimulus? Did they see feature X? A theory that predicts that I am experiencing a red square when I am convinced that I am seeing a blue triangle has serious problems. Reporting about our introspection is at the very core of consciousness science, and is a rich source of scientific data. Crucially, the subjective reports based on introspection used in consciousness science are not enough to evade the UA. Indeed, everyone agrees that subjective reports are third-person data and therefore are subject to the UA. That is to say, when a subject reports what their first-person experience is like, we know logically that the causal structure is neither necessary nor sufficient to explain the report.

In contrast to (reported) introspection, "*pure* first-person experience" denotes any putative kind of experience that *can only be directly accessed by the subject* (Chalmers, 1996; Cohen & Dennett, 2011). These experiences are inherently private, because as soon as we report about first-person experiences, they become third-person data. Since pure first-person experiences cannot be shared without becoming third-person data, pure first-person experiences by themselves are not scientific data, and therefore cannot support or disprove scientific theories - only reports about them can do that. See also Chalmers (1996), Cohen & Dennett (2011), Frankish (2016) and Goff (2019) for in depth arguments.

Since introspective reports are third-person data and therefore subject to the UA, IIT needs to find support in such *purely* first-person experiences. But *pure* first-person experiences *cannot be used as scientific data.* Addressing first-person experiences scientifically requires translating them into third-person data, and the UA applies. For example, using psychophysics to test if IIT's axioms are true for humans is an interesting scientific endeavor. But as soon as we do this, the axioms become third-person data measured by i/o experiments, and are no longer the pure first-person experience needed to rescue IIT.

In summary, IIT cannot evade the UA by relying on introspection, because this would require *purely* first-person data, detached from third-person data. But this kind of data doesn't exist. Data need to be communicable – and purely first-person experiences are not.

2.2. (Neo)-Cartesian reasoning is unscientific

Still, IIT proponents argue that IIT's axioms hold undeniable ground truth, *grounded* in pure first-person experience (Negro, 2020; Oizumi et al., 2014). Oizumi et al. (2014) explain this by analogy to Descartes' famous cogito ergo sum. One can doubt all aspects of the external world. For example, a malicious demon may fool us into perceiving things that don't exist, similarly to hallucinations. However, one cannot doubt one's own existence without ending up in a contradiction. In a similar form of Cartesian reasoning, pure first-person experience can give rise to ground truth. There is no need for sharable third-person data.

We think that such (neo-)Cartesian reasoning is incompatible with the scientific method. No scientist can claim to have ground truth independent of data. The task of scientists is to confront theoretical claims with data. This is how theories are compared and science progresses. No other empirical science is based on the assumption that we have ground truth, and consciousness science should be no exception.

In addition, even if we made an exception for consciousness science and based it on purely first-person ground truths, these ground truths would at least need to be self-evident and uncontroversial. But the axioms are neither self-evident nor uncontroversial. If they were, there would be no debate about them. But there *is* debate. Indeed, some people explicitly argue against IIT's axioms (Bayne, 2018) and some have theories of consciousness that directly contradict some of IIT's axioms (Frankish, 2016; Zeki, 2007). Is consciousness necessarily unified? Maybe for most of us it is, but maybe not for a patient with Balint's syndrome. How to decide? There is no way to resolve this clash through pure first-person experience. That is where empirical science (based on third-person data) becomes necessary and the UA applies.

Lastly, even if we agreed that consciousness science is the only science based on purely first-person ground truths, and even if we all agreed on the axioms, this would still not be enough. Indeed, how to translate the axioms into Φ and qualia space is obviously not given by first-person experience. This is problematic since, as Cerullo (2015) and Hanson and Walker (2021) have shown, there are many ways to deduce Φ from the axioms (see also Aaronson, 2014a). The current version of IIT is just one out of many and no pure first-person experience can justify this choice. Again, third-person data are needed.

In summary, IIT's axioms cannot appeal to purely first-person ground truths.

2.3. Third-person experiments cannot support IIT as a theory of consciousness

If purely first-person experience cannot provide a solid foundation for IIT, what about the third-person experiments claimed to support IIT? For example, a measure of brain activity inspired by IIT, called the Perturbational Complexity Index (PCI), is able to robustly determine the conscious states of patients (although this measure is also compatible with other theories, such as Global Workspace Theory; Mashour et al., 2020). As another example, awake drosophila flies seem to have a higher Φ than anaesthetized ones (Leung et al., 2021). A third example is the ongoing adversarial collaboration testing IIT vs. GWT using psychophysical and neuroimaging techniques (Melloni et al., 2021). What do we make of these experimental achievements?

According to the UA, these experiments in fact do not support IIT *as a theory of consciousness.* Therefore, there must be another reason why PCI and other approximations of Φ seem to correlate with consciousness in humans (and fruit flies). One possibility is that these measures index something else, like representational capacity or processing efficiency (Mediano et al., 2021; Merker et al., 2021; Toker & Sommer, 2019), that happens to be tightly coupled with human consciousness and/or is a necessary condition for consciousness, but is not identical to it. For example, having brain damage due to head trauma or a stroke probably does not *selectively* abolish consciousness, leaving processing efficiency and other functions intact. So, a reliable general measure of processing efficiency might work out to be a good measure of consciousness because the two tend to go hand in hand.

In other words, IIT may be a good theory of perception, for example: low Φ reflects poor efficiency and high Φ reflects good efficiency, without being a theory of consciousness *per se*. This assertion is perfectly testable in principle. Indeed, we could create good independent measures of efficiency that we then can test against Φ. For example, we may count the number of neurons needed to implement a function, and test whether systems with higher Φ require fewer neurons or less energy to implement that function. In this way, IIT may tell us that photodiodes implement a similar kind of strategy for efficient processing than awake brains, which would be a valuable insight.

The unfolding argument is based on the fact that, unlike efficiency, for which we have good measures independent of i/o functions, we have no scientific measure of consciousness independent of i/o functions. Therefore, no experiment can support the idea that the photodiode is or is not *conscious* in virtue of its causal structure, because causal structure is dissociated from empirical results about consciousness. For this reason, IIT cannot be tested as an empirical theory of *consciousness*.

2.4. Dissociative epiphenomenalism

In the following, we will argue that IIT implies that consciousness is epiphenomenal, i.e., it has no influence on any type of behavior including motor actions, verbal responses, etc. Not only is the magnitude of IIT's consciousness epiphenomenal but also the content of IIT's consciousness is fully dissociated from behavior. As a consequence, IIT implies that creatures may exist that experience things totally different from what they report (see also Michel, 2021). Hence, if IIT is true, we cannot trust reports about first-person experience and, therefore, these reports about first-person experience cannot serve as a foundation for IIT's axioms. In this way, IIT undermines its own foundations.

As mentioned, IIT implies that two creatures with identical i/o functions can have different consciousnesses due to different causal structures (Oizumi et al., 2014). Hence, the two creatures show identical behavior in all respects. For example, all experimental

stimulus-response mappings are the same, i.e., the very same stimuli lead to the same actions in the two creatures. As a consequence, IIT entails that consciousness is completely epiphenomenal. The very same actions can occur whether or not consciousness is involved.

This argument holds true not only for the binary conscious vs. unconscious distinction but for all levels of consciousness. Consider a robot with its i/o function and associated Φ. We can change its wiring such that the i/o function stays the same but the Φ varies freely. In fact, we can create infinitely many such robots with different sets of Φs. Hence, IIT predicts that the level of consciousness, determined by Φ, is fully dissociated from behavior. This dissociation is a straightforward consequence of the dissociation of causal structure from i/o functions. If causal structure determines consciousness independently of i/o functions, then, by definition, consciousness cannot matter for behavior, i.e., i/o functions.

Further, IIT's consciousness implies, as we call it, a form of *dissociative epiphenomenalism*, in which not only the *magnitude* of consciousness, determined by Φ, is epiphenomenal and fully detached from i/o functions, but also its *content*. We have shown (Appendix in Doerig et al., 2019) that we can construct robots that never experience what they report to perceive, feel, and think. This can always be achieved by changing the causal structure without changing i/o functions. For example, we can change parts of the feedforward wiring of the unconscious feedforward robot to a recurrent network without changing the i/o function. Since this recurrent network has the highest Φ, IIT predicts that the robot will be conscious and experience the constellation in qualia space corresponding to the recurrent network. Hence, we can change the recurrent network to create any content we want without changing i/o functions, i.e., we can make the robot consciously perceive whatever we want.

This has extremely strange implications. For example, we may wire this robot to always experience the smell of coffee, independently of what it claims to perceive. The robot may perform complex tasks and report about them, but, according to IIT, consciously never experiences anything other than the smell of coffee. Alternatively, we can wire up the robot so that conscious percepts permanently change, completely independently from the robot's behavior (for example, the robot may always report that it is smelling coffee, but according to IIT its experiences are ever-changing). Of course, IIT does not claim that such creatures actually exist, but it is a direct implication of IIT. In addition, creating such systems is straightforward (Hanson & Walker, 2020).

Since IIT's consciousness is truly dissociated from i/o functions, *we may all be* such creatures. We may in truth all be perpetually experiencing the smell of coffee but be unable to express it, because neither the level nor the content of consciousness has any impact on reports or any other third-person data. You may conclude, based on pure first-person experience, that this scenario does not hold true for you. But you cannot share this pure first-person experience. You can only share third-person data that is dissociated from your (IIT-)consciousness. We may all verbally agree that the scenario is wrong but, as just shown, third-person verbal agreement is no evidence about our true conscious states (according to IIT) because consciousness is fully dissociated from what we can behaviorally express. Thus, we can never know whether a report truly reflects the content of (IIT-)consciousness.

In other words, IIT *entails* that behavior is fully dissociated from consciousness. We cannot trust reports about first person experience. Consequently, IIT *implies* that (pure) first-person experience cannot be communicated and therefore cannot be a foundation for its own reasoning. Accepting IIT based on the putative first-person truth of its axioms leads us to reject the axioms. This self-contradiction and the other problems mentioned earlier all stem from the fact that causal structure is completely dissociated from behavior.

3. Discussion

IIT starts from ground truth "axioms" and derives postulates, as in mathematics. The initial hypothesis is that, when it comes to consciousness, it matters more that two systems are similar in terms of their causal structure than in the functions they implement. Consciousness does not depend on what systems *do,* but rather on *how they do it.* This hypothesis is a valid starting point, but, like any other scientific hypothesis, it requires empirical justification. The UA claims that this cannot be done.

To summarize the debate, the UA shows that there can be systems with identical i/o function but different causal structures (e.g., different Φs) and hence different consciousness according to IIT, and vice-versa. For example, the two robots that we described earlier are identical in everything we can measure or find out from i/o experiments, but only one is conscious. From a standard scientific viewpoint based on third-person data, this means that IIT is false because it predicts different consciousness for systems with the same observable data about consciousness. This part of the argument is well accepted, acknowledged by both sides of the debate, including IIT proponents (Albantakis, 2020; Kleiner, 2020; Kleiner & Hoel, 2021; Oizumi et al., 2014). The only observable difference between a recurrent and an unfolded system is the causal structure itself, but we argued that justifying a theory based only on causal structure is circular.

This circularity was strongly opposed by proponents of IIT stating that the differences in the causal structure do matter for consciousness. Negro (2020) and Kleiner (2020) argue that whereas indeed relying directly on Φ is circular, Φ itself is logically derived from axioms rooted in first-person experience, providing ground truth. However, we have argued here that this requires appealing to *pure* first-person experience independent of third-person data. Pure first-person experiences are not scientific data. They cannot be used as ground truths. In addition, as we have shown, third-person data cannot support IIT as a theory of consciousness either. Furthermore, even if one bluntly assumes that IIT is true, we still end up with contradictions. Since IIT's level and content of consciousness are fully dissociated from any behavioral i/o function, IIT's consciousness is acausal, without any impact on the world. We may consciously perceive things and report completely independent things. Therefore, accepting the implications of IIT leads us to reject reports as a source of data about consciousness. But IIT's axioms rely on reports about pure first-person experiences, closing the contradictory loop. Cohen and Dennett (2011) used a related argument against the notion of purely phenomenal consciousness without access consciousness.

Our argument that IIT leads to epiphenomenalism is similar to the "zombie argument for physicalism" (Carroll, 2021; see also

Balog, 1999; Brown, 2010; Frankish, 2007). The idea is that, if a theory admits the possibility of identical physical duplicates without consciousness (i.e., zombies), then the theory faces the following challenge: since you and your zombie have all the same observable properties, the theory in fact explains nothing – it adds nothing to our understanding of the world. For example, it does not explain why you behave as you do in a masking experiment, why you may (or may not) think there is a hard problem or why you are interested in (or irritated by) the current paper. Despite these similarities, an important difference between the UA and the zombie argument for physicalism is that the UA is not based on a thought experiment and does not require exact physical duplicates or other possible worlds. It is a mathematical fact that different systems can implement the same function with different causal structures in our world.

Usher (2021) has attacked the UA by claiming that certain aspects of recurrent networks cannot be approximated in a feedforward fashion. We agree that, if differences in causal structure always lead to observable differences in i/o functions, then the UA is weakened. In this respect, one might be troubled by the fact that recurrent networks are Turing Complete (Siegelmann & Sontag, 1995), while feedforward networks are not. However, first, this result is proven using discretely unrolled recurrent networks, which can also be seen as feedforward. Second, more importantly, we focused on comparing feedforward and recurrent networks only for the sake of simplicity. The very same arguments apply when comparing different recurrent networks and many other algorithms. As mentioned (Doerig et al., 2019), there are many more varieties of universal function approximators, many of which are also Turing Complete, such as different varieties of recurrent networks, transformers and neural GPUs (i.e., modern neural network architectures; Pérez et al., 2019), cellular automata (Cook, 2004), many programming languages, and computers. Even Powerpoint (Wildenhain, 2017) and Magic: The Gathering (Churchill et al., 2019) are Turing Complete. Each of these systems is as computationally flexible as a recurrent network and has a completely different causal structure. Hence, the UA holds: any function, and even every process can be implemented with many different causal structures (see also Kleiner & Hoel, 2021; and DR5 for further discussion).

Importantly, we are not i/o chauvinists. We do not argue that i/o functions define consciousness at all. We are not saying that causal structure does not matter just because we *assume* that consciousness is a function. We do not assume anything metaphysical. The UA only shows that IIT is false or circular. Hence, classic attacks against functionalism, such as the zombie and blockhead arguments (see DR3) do not make the unfolding argument any less valid.

An important question is whether the UA shows that causal structure and experimental results about consciousness *must* be dissociated, or whether they merely *could* be. Does the UA apply only to some weird exotic robots that don't actually exist? No. The UA applies to all systems. As mentioned in the "falsification" section, the point of the UA is not to provide a "surprising" counter-example that we build and suddenly realize "oh, this feedforward robot is unconscious, so it falsifies causal structure theories". Rather, the UA logically shows that, when we observe an empirical result about consciousness*, it is provably not due to causal structure.* Of course, in certain systems, reports and causal structure do co-occur. For example, in humans, causal structure seems to track reports about consciousness in several cases. Still, as we argued, this cannot support IIT as a theory of *consciousness.* The causal structure is neither necessary nor sufficient for any report about consciousness. The UA shows that this *is* the case, not merely that it *could be* the case. To avoid this problem, IIT requires another axiom about some kind of parallelism, or a mystical force, enforcing that reports about pure first-person experiences do in fact match these pure first-person experiences.

To be clear, we are not denying the importance of the brain for consciousness research. Quite the contrary, brain research will guide us to hypotheses about consciousness, which pure first-person experiences cannot deliver. For example, recent research has shown that conscious perception of visual stimuli correlates with a certain type of processing in the dendrites of pyramidal cells in layer 5 (as proposed a long time ago by Crick and Koch, 2003), which enables gating of information (Aru et al., 2020). Hence, the *functional* aspect of gating may be important for consciousness, mediated by certain types of causal structure. Different causal structures leading to the same gating should leave consciousness unaffected. As another example, there is currently a debate about the role of prefrontal cortex for consciousness (Michel & Morales, 2020). Resolving this debate empirically will help us understand which functions of the prefrontal cortex are linked (or not) to consciousness, thereby shedding light on which processes are important for consciousness. Whatever may come of it, it is important that brain research serves as a generator for hypotheses and not as an a-priori truth-maker.

4. Conclusions

Amongst current theories of consciousness, IIT is arguably the most precise – making quantitative predictions about both the level and state of consciousness of any given system. This rigor should be lauded, especially in the current context where the vagueness of most theories makes comparisons difficult (Doerig, Schurger, et al., 2020). Despite these qualities, IIT has been strongly criticized in recent years (Aaronson, 2014b, 2014a; Cerullo, 2015; Doerig et al., 2019, Doerig, Schurger, et al., 2020; Hanson, 2021; Hanson & Walker, 2020; Hanson & Walker, 2021; Lau & Michel, 2019; Merker et al., 2021). We think that one of the main problems of IIT is that it takes (contingently) necessary aspects of consciousness, such as integration, to be sufficient for consciousness (as do other theories in the field; Doerig, Schurger, et al., 2020). This leads, among other things, to inactive XOR grids that are more conscious than humans (Aaronson, 2014b) and to the UA. A common thread in IIT's response to these criticisms has been that we must rely more on first-person experience than on third-person data. Here, we argued that this approach is fundamentally flawed.

To summarize, given standard scientific criteria restricted to i/o experiments, IIT is false because it entails that systems with identical i/o functions can have different consciousness. Defining consciousness in terms of causal structure cannot remedy the situation because it is circular and hence leads to unfalsifiability. Going beyond standard science and appealing to pure first-person experience cannot rescue IIT either and ultimately leads to dissociative epiphenomenalism and profound contradictions. Hence, causal structure theories cannot use pure first-person experience to avoid the UA. In general, theories that entail a complete dissociation of consciousness and i/o functions cannot be tested empirically and have extremely strange and self-contradictory consequences.

Author statement

This ms is not under considerations with another journal. There are no experiments and hence not ethics approval necessary.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Michael Herzog was supported by the Swiss National Science Foundation (SNF) grant "Basics of visual processing: from elements to figures". Adrien Doerig was supported by the SNF grant n.191718 "Towards machines that see like us: human eye movements for robust deep recurrent neural networks".

Detailed Replies

DR1. Robots, brains, and biological chauvinism: reply to Albantakis (2020). Albantakis (2020) proposed that the robot example makes no sense -at the moment- because we do not know whether robots have consciousness at all and hence one cannot use robots for falsification. On the other hand, we know by first-person experience that we are conscious. Therefore, we should trust human reports but not anything else, so the robot example does not falsify IIT. Likewise, Albantakis proposes that we should not trust the reports of humans with implants that change their brain's causal structure (see also https://youtu.be/mWl-gW75O94). In general, the argument is that we should only trust reports of systems that are close enough to ourselves, because the only first-person experience we have is from ourselves. However, this argument is problematic for four reasons.

First, we would like to reiterate that the term "falsified", which we used in our previous contribution, may have caused confusion. We have shown that IIT is logically *false*, rather than falsified, or circular (given standard scientific criteria). There is no need to do experiments, neither in robots nor humans. IIT is *logically* flawed and thus the above argument simply does not apply. The only way out is via pure first-person experience, which leads to circularity and/or epiphenomenalism, as argued above.

Second, it is important to note that the UA does not only apply on the purely behavioral i/o level. It applies within systems as well. For example, let us assume a person has a conscious experience of X with a certain causal structure. The UA guarantees that we can replace tiny parts of the brain's causal structure and thus change Φ without changing the "behavioral" i/o function. For example, we may replace just a few feedforward neurons in IIT's main complex determining the percept, by even fewer recurrent neurons while keeping all the other neural and behavioral i/o functions constant. All neural i/o functions to and from the tiny brain part we changed stay the same. Should we not trust the reports of such a person? If not, why not? In addition, we do not need to surgically change the brain's wiring. For example, the brains of stroke patients or patients with hydranencephaly have strong differences compared to "healthy" brains – much stronger than the slightly modified brains we just mentioned. Should we not trust their reports about consciousness? Moreover, the brain changes itself perpetually and often drastically. Should we not trust our own reports from the past because our brains have changed (after all, we do not have any direct first-person evidence that we *were* conscious a year ago)?

Third, without a clear criterion delineating which systems should be trusted as sources of data for consciousness, IIT is free to choose the cases that suit it. For example, a recent study showing that Φ is higher in awake than anaesthetized fruit flies is often cited as evidence for IIT (Leung et al., 2021). However, fruit fly brains are much, much more different from "normal" human brains than the examples given above or than a human brain with a small implant changing the brain's causal structure. If IIT wants to evade the UA because it does not trust the reports of modified systems, it should discard fruit fly studies as irrelevant. One cannot just pick the cases favorable to one's theory.

Fourth, even using "normal" human reports as empirical data is problematic for IIT. Proponents of IIT may argue that we can trust the reports of humans with small brain changes or a stroke, but that these changes come with slight changes in consciousness. However, there is a catch: IIT needs to postulate that brain changes that keep i/o functions identical always keep Φ constant too. Indeed, if such changes could happen then we could no longer trust the reports of humans about their own consciousness, because their causal structure may change (therefore changing consciousness) without changing their reports! The mathematical theorems underlying the UA guarantee that there are abundantly many possibilities for this to occur, making it hard to imagine why such a replacement cannot happen naturally. It seems that IIT needs another axiom to rule this out- which does not come from first-person experience.

Relatedly, one may propose to only trust reports of biological systems, embracing "biological chauvinism", the idea that only biological systems can have consciousness. However, this proposition runs into the same problems: we can replace a brain region with a non-biological implant that has the same i/o function as the biological tissue, thus completely dissociating "biologicalness" and i/o functions. This exposes biological chauvinism to the same problems as causal structure theories.

DR2. Reply to Tsuchiya et al. (2020). Tsuchiya et al. (2020) argue that the UA relies on a new mode of methodological behaviorism. But labelling us behaviorists only distracts from the more substantive points that we have tried to make. It also could not be farther from the truth, as discussed in the introduction. In fact, what they call behaviorism is what we call empirical science. Science depends on publicly observable physical measurements: we measure physical phenomena and mechanisms and try to understand how they work.

Tsuchiya et al. go on to assert that our stance effectively focuses only on the publicly available data, such as reports *about* consciousness. Our stance indeed focuses on things that we can measure, which include behavior, but also brain activity (as long as it is not identified with causal structure a priori) and other kinds of physiological measurements. This is the standard stance in empirical research, including research on consciousness. To wit, in a recent survey, Francken et al. (2021; Fig. 4) asked 232 consciousness researchers about approaches to study consciousness. *All* answers are compatible with this stance and focus on publicly observable data. *There are no non-publicly-observable data in science* (indeed, how could you write a journal article or give a talk that includes non-publicly observable, non communicable data). According to Tsuchiya et al., a consciousness researcher *must* be imputing some non-observable properties to their subjects. But then what, precisely, do they mean by "non-observable"? Are they referring to a ghost in the machine, which is what non-observable normally refers to? Even the equation-inspired Higgs boson and dark matter are subject to the scientific standard of measurement. If it exists, then we ought, in theory, be able to measure it publicly. That includes consciousness, if we want to remain in the realm of science.

In reference to the first premise of the UA, that in science we rely on physical measurements, Tsuchiya et al. remark that they agree "in a limited way". However, there can be no slack here. Relying on physical measurements is a hard rule of science – if you bend that rule then it is no longer science. Tsuchiya et al. maintain that the 'ground truth' data in consciousness science are one's own conscious experiences. We have shown above the problems that this leads to. By contrast, we maintain that the ground truth data in consciousness research are reports about subjective experience (including our own). If one's own purely first-person *experiences* are the ground truth for consciousness science, then you had best keep them to yourself. Because the moment you tell someone else about them, these become publicly available reports. We encourage the reader to try the thought experiment of stepping, carefully, through what it would take to *scientifically* study the state or content of consciousness entirely on your own, using "pure first-person experience" (with or without the help of equipment). Taking this exercise seriously shows that it cannot be done – peer review would be problematic, to say the least.

Tsuchiya et al. advocate for variants of our first premises, and argue that their modified premises most closely resemble the science of consciousness as it is currently practiced. The first of their new premises, which they call CP1, proposes that a consciousness scientist should base her theory on multiple modes of evidence *always* with respect to the link with conscious experience itself. But then that is no different from the approach that we advocate, except that you just need to insert "reports about" before "conscious experience". The crucial question, according to Tsuchiya et al., is whether or not this "input-output function" includes the *generation of conscious experience*. But then what precisely does this mean? Does it refer to the generation of something non-measurable? If so, no science can be done about it. If not, then what, precisely, does it refer to, scientifically?

Regarding the second premise of our UA (P2) Tsuchiya et al. contend that it should be rewritten to read "For any recurrent system with a given input-output function (*excepting inputs-to or outputs-from a conscious experience*), there exist feedforward systems with the same input-output function (and vice-versa)" (our emphasis). We never made such an exception, and do not advocate for it. Besides requiring non-scientific pure first-person data for the reasons outlined in this article, this statement begs the question of what these authors mean by "conscious experience" and why we have to exclude inputs to it or outputs from it. Do they mean that we have to exclude those because conscious experience is not measurable? This brings us back to our original argument, which we stand behind, that non-measurable things are not the stuff of empirical science.

Finally, Tsuchiya et al. argue that consciousness science should search for an isomorphism between physical and phenomenal structures. However, this amounts to nothing more than identifying the neural correlates of consciousness, characterizing or describing (the content of) consciousness and its attendant brain activity. This is a classic and useful endeavor, but is well known to fall short of *explaining* consciousness. Tsuchiya et al. propose that their isomorphism approach requires "inventing no-report paradigms". No-report paradigms (Tsuchiya et al., 2015) indeed help to cope with some important confounds of experimental setups, but fail to address other confounds (e.g. Peters & Lau, 2015). Importantly, they still ultimately depend on reports, as outlined in the introduction. We agree that new paradigms that address different confounds are important, but they are not a magic ingredient that readily leads to a true explanation of consciousness, which is what a theory is supposed to offer.

DR3. Blockheads, Lakatos, and consciousness. Negro defends IIT with a Lakatosian perspective of science, according to which only certain claims need to be relaxed if a theory is falsified, in order to protect the core claims. However, any "amendment" to the axioms invalidates them as a viable source of ground truth and the UA resurfaces. One cannot have their cake and eat it too. If first-person experience does not really provide undeniable ground truth, then we are back to square one: third-person empirical tests are needed to support the axioms and the UA applies. Hence, there is no rescuing IIT from the UA, because pure first-person experience either needs to be unscientifically trusted in a neo-Cartesian leap of faith, or third-person data is needed. To be clear, we do *not* claim that all of IIT's axioms are *wrong*. But they cannot be considered *ground truths* without (third-person) empirical support.

Negro proposed to support the notion of first-person ground truth by analogy to Block's Blockhead argument (Block, 1981). The Blockhead is a lookup table that plays chess identically to a Grand Master and, thus, has the same i/o function for chess (***f***(chess board situation) -> move, etc.). However, arguably, it does not have intelligence. Hence, system X (the Grand Master) can have a different property Z (intelligence) from system Y (the Blockhead), even though their i/o functions are identical. Likewise, according to Negro, feedforward and recurrent networks can differ in consciousness despite having the same i/o function. We have certain reservations about this reasoning (for example, the Blockhead replicates the i/o function of *chess,* not *intelligence* – so no wonder it is not intelligent), but they are not crucial here. The main problem with this argument is that, although there are potentially measurable reasons to believe that the Blockhead is not intelligent (e.g., it could not understand how to play if we invert the roles of the knight and bishop), there is *no reason* why the recurrent robot has consciousness and the feedforward robot does not - unless one believes in the a-priori ground truth of IIT's axioms, leading to the problems described above.

Here is another way to phrase the above point. The blockhead argument shows that one can play chess in two different modes,

without and with intelligence. Here, intelligence simply describes a kind of processing. For example, a look up table provides all information explicitly whereas an intelligent system perhaps uses combinations of abstract rules, which can be generalized easily. We can agree or debate whether only the latter is "intelligent", but the point is that this is a debate about different kinds of processing. Similarly, systems can compute information in a feedforward or recurrent manner. However, in the case of consciousness, there is something additional to explain, namely, subjective experience. For example, why do we consciously perceive stimuli only in certain cases? Hence, this is not only a debate about different kinds of processing as for intelligence. It is about which kinds of processing *lead to conscious percepts*. The latter has no analogy in the blockhead argument.

DR4. Does the UA rule out most theories of consciousness? Kleiner (2020) and Kleiner & Hoel (2021), using a *reductio ad absurdum* type of argument, argue that a generalization of the UA rules out most theories of consciousness, not only causal structure theories. Hence, they propose that there must be something wrong with the UA, which they suggest is premise P3. Kleiner points out that a giant look-up table (or other "trivial" systems) can realize any i/o function. For example, he argues that the global workspace of the Global Neural Workspace Theory (GNWT) can be replaced by a look up table without changing i/o functions. Hence, if GNWT assumes that the lookup table is not conscious, it runs into the same problem as IIT: consciousness does not depend on i/o functions. We agree with this argumentation if the global workspace is defined in terms of causal structures, as Kleiner seems to suggest. However, if the workspace is defined in functional terms, then the lookup table also realizes a global workspace. Contrary to causal structure, there is no mathematical theorem stating that the same i/o functions can be realized with and without a global workspace (see also Ganesh, 2020). In summary, we agree that Kleiner's argument applies to theories that *identify* consciousness with a certain non-functional process *claimed to be necessary and sufficient* (as, indeed, many theories do (Doerig, Schurger, et al., 2020)). However, theories may be cast in functional terms, or propose that consciousness should not simply be identified with a single process, just as life is not identified with a single process (Machery, 2012).

In summary, unlike Kleiner and Hoel, we argue that there is room for consciousness research without abandoning P3. Indeed, while the UA provides *logical guarantees* that consciousness and i/o functions are doubly dissociated according to IIT and other CSTs, there is no such theorem for non-causal structure theories. We agree there is a challenge for consciousness research, but it is not worse than the well-known challenges linked to the hard problem, which have not stopped consciousness science.

DR5. Recurrent networks cannot be unfolded. Usher (2021) argues that the mathematical theorems used in the unfolding argument do not in fact entail that feedforward networks can be functionally equivalent to recurrent networks in dynamical cases, for example involving temporally extended stimuli. There are two main arguments.

First, Usher agrees that unrolling a recurrent network through time leads to a functionally equivalent feedforward network. Unrolling through time is a common technique used in machine learning to train "recurrent networks" (see e.g., LeCun et al., 2015). Since computer operations are sequential, they cannot truly implement recurrence. To solve this problem, unrolling through time copies the whole network for each timestep, and treats timesteps in sequence: computations for timestep 1 are carried out, fed as input for timestep 2, etc. Usher accepts unrolling in general but dismisses unrolled networks in the context of the UA, because they are too large. For example, simulating 1 second of activity with a 10 ms resolution requires copying the network 100 times.

If consciousness necessarily involves the entire brain with Plank-scale temporal accuracy, the unrolled networks would be very large indeed. However, these are practical questions and not principled ones. Dismissing unrolled networks is not warranted in the context of the UA. First, the UA applies not only to entire human brains, but also to sub-parts of the brain (see DR1 and Doerig et al., 2019), and to smaller systems easily implementable with unrolled networks, such as our robots example or the examples of Hanson and Walker (2020). Second, in practice, large scale unrolled networks can be used to model human reaction times (Spoerer et al., 2020) and ventral stream dynamics (Kietzmann et al., 2019). The latter network simulates 0.6 s with a 10 ms resolution, similarly to the example deemed "too large" by Usher. This suggests that many neural computations relevant for processing and consciousness across large parts of the brain can in fact be unrolled even in practice. Third, even though it is true that recurrent networks can implement certain functions, especially temporally extended ones, much more efficiently than feedforward networks, this in no way supports the theory that recurrence is identical to consciousness. As we explained in the section "Third-person experiments cannot support IIT as a theory of consciousness", causal structures are good candidates for theories of efficient processing, but cannot be supported as theories of *consciousness.* Likewise, in Doerig et al. (2019), we argued that evolution requires efficiency, which explains why the brain is recurrent – but does not imply that recurrence is identical to consciousness.

Having dismissed unrolled networks, Usher proposes that non-unrolled feedforward networks behave differently from recurrent networks when dynamics are involved. To support this, Usher gives examples where feedforward and recurrent networks behave identically on static inputs, but differently on temporally extended inputs.

However, this misses the point of the UA. The UA is based on the fact that, *given a function*, both recurrent and feedforward inputs can implement that function. For example, the function mapping binocular rivalry stimuli to subject responses can be implemented by recurrent and feedforward networks. This is different from saying that *given two network that behave similarly on task X,* these networks are identical in all other respects. In all the examples given by Usher, two networks behave similarly on a static task and differently on a dynamic task. What this shows is that *the networks implement a different function*. This in no way suggests any link with consciousness. It is like saying: "You might have thought that the networks implemented the same function, but you were wrong – in fact we notice the functions are different when time comes into play!". We know from mathematics that there are networks with different causal structure that are also functionally identical under dynamical settings, but we have just shown that this is not one of them.

Similarly, Usher explains that recurrent networks and feedforward networks behave differently under ablations: cutting a recurrent connection is not the same as cutting a feedforward one. But, again, this misses the point of the UA. The UA cares about cases where the recurrent and feedforward networks implement the same function. But after the ablation, they *don't implement the same function anymore,* as Usher shows by pointing out that their behaviour changes. There still is a feedforward network with the same function as

the ablated network, but it is a different one.

We would like to stress that there *are* feedforward networks that implement any function, even when time plays a role. First, unrolled networks are used by researchers using machine learning every day to implement temporally extended computations, including for large-scale brain modelling. Second, there are many ways of processing a temporally extended stimulus in feedforward networks. As Usher mentions, one way is to inject later stimuli in deeper layers, but there are many different ways to encode time. Third, modern day computers are perfect examples of systems that can obviously implement complex temporal functions with $\Phi = 0$. In fact, even the "recurrent" examples given by Usher were run on a feedforward computer! Modern day computer technologies offer a dazzling display of the versatility of feedforward computation. More generally, as mentioned in the main text, focussing on feedforward networks is a simplification. Many systems, all with different causal structures, are equally computationally versatile. Recurrence is powerful, but it is not unique.

In summary, Usher's argument misses the main point by focusing on cases where the feedforward and recurrent networks do *not* in fact implement the same i/o function. The UA is based on cases in which systems with different causal structures *do* implement the same function. These equivalent systems exist, as proven by mathematical theorems and numerous practical applications. Recurrent processing is of course a very important topic (which we study ourselves, e.g., Doerig, Schmittwilken, et al., 2020), and has different strengths and weaknesses compared to feedforward computing. But there is no evidence suggesting recurrence is a magic ingredient that creates consciousness.

References

Aaronson, S., 2014a. *Giulio Tononi and me: A phi-nal exchange*. http://www.scottaaronson.com/blog/?p=1823.

Aaronson, S. (2014b). *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. https://www.scottaaronson.com/blog/?p=1799.

Albantakis, L. (2020, September 14). *Unfolding the Substitution Argument*. Conscious(Ness) Realist. https://consciousnessrealist.com/unfolding-argument-commentary/.

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular mechanisms of conscious processing. *Trends in Cognitive Sciences*.

Balduzzi, D., & Tononi, G. (2009). Qualia: The geometry of integrated information. *PLoS Computational Biology, 5*(8), Article e1000462.

Balog, K. (1999). Conceivability, possibility, and the mind-body problem. *Philosophical Review*, 497–528.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness, 2018*(1), niy007.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review, 90*(1), 5–43.

Brown, R. (2010). Deprioritizing the a priori arguments against physicalism. *Journal of Consciousness Studies, 17*(3–4), 47–69.

Carroll, S. (2021). Consciousness and the Laws of Physics. *Journal of Consciousness Studies, 28*(9–10), 16–31.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., & Tononi, G. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine, 5*(198), 198ra105-198ra105.

Cerullo, M. A. (2015). The problem with phi: A critique of integrated information theory. *PLOS Computational Biology, 11*(9), Article e1004286.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Churchill, A., Biderman, S., & Herrick, A. (2019). Magic: The Gathering is Turing Complete. *ArXiv:1904.09828 [Cs]*. http://arxiv.org/abs/1904.09828.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences, 15*(8), 358–364.

Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems, 15*(1), 1–40.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*(2), 119–126.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492.

Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., & Voss, H. U. (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science Advances, 5*(2), eaat7603.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology, 16*(7), Article e1008017.

Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, 1–22. https://doi.org/10.1080/17588928.2020.1772214

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

Francken, J., Beerendonk, L., Molenaar, D., Fahrenfort, J., Kiverstein, J., Seth, A., & van Gaal, S. (2021). *An academic survey on theoretical foundations, common assumptions and the current state of the field of consciousness science*.

Frankish, K. (2007). The anti-zombie argument. *The Philosophical Quarterly, 57*(229), 650–666.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11–12), 11–39.

Ganesh, N. (2020). C-Wars: The Unfolding Argument Strikes Back—A Reply to'Falsification & Consciousness'. *ArXiv Preprint ArXiv:2006.13664*.

Goff, P. (2019). *Galileo's error: Foundations for a new science of consciousness*. Vintage.

Graham, G. (2000). *Behaviorism*.

Hanson, J. R. (2021). *Falsification of the Integrated Information Theory of Consciousness*. Arizona State University.

Hanson, J. R., & Walker, S. I. (2019). Integrated information theory and isomorphic feed-forward philosophical zombies. *Entropy, 21*(11), 1073.

Hanson, J. R., & Walker, S. I. (2020). Formalizing Falsification of Causal Structure Theories for Consciousness Across Computational Hierarchies. *ArXiv Preprint ArXiv: 2006.07390*.

Hanson, J., & Walker, S. I. (2021). On the Non-uniqueness Problem in Integrated Information Theory. *BioRxiv*.

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy, 21*(12), 1160.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Kent, L., & Wittmann, M. (2021). *Time Consciousness: The Missing Link in Theories of Consciousness*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, 116*(43), 21854–21863.

Kleiner, J. (2020). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition, 85*, Article 102981.

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness, 2021*(1), niab001.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494–501.

Lau, H., & Michel, M. (2019). *On the dangers of conflating strong and weak versions of a theory of consciousness*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leung, A., Cohen, D., Van Swinderen, B., & Tsuchiya, N. (2021). Integrated information structure collapses with anesthetic loss of conscious arousal in Drosophila melanogaster. *PLoS Computational Biology, 17*(2), Article e1008722.

Machery, E. (2012). Why I stopped worrying about the definition of life... And why you should as well. *Synthese, 185*(1), 145–164.
Mallatt, J. (2021). A traditional scientific perspective on the integrated information theory of consciousness. *Entropy, 23*(6), 650.
Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron, 105*(5), 776–798.
Mediano, P. A. M., Rosas, F. E., Farah, J. C., Shanahan, M., Bor, D., & Barrett, A. B. (2021). Integrated information as a common signature of dynamical and information-processing complexity. *ArXiv:2106.10211 [Nlin, q-Bio]*. http://arxiv.org/abs/2106.10211.
Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science, 372*(6545), 911–912.
Merker, B., Williford, K., & Rudrauf, D. (2021). The Integrated Information Theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences*, 1–72.
Michel, M. (2021). If IIT is true, IIT is false (The Unfolded-Tononi Paradox). https://matthiasmichel.blogspot.com/2021/10/if-iit-is-true-iit-is-false-unfolded.html.
Michel, M., & Morales, J. (2020). Minority reports: Consciousness and the prefrontal cortex. *Mind & Language, 35*(4), 493–513.
Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: The case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences, 19*(5), 979–996.
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), Article e1003588.
Pérez, J., Marinković, J., & Barceló, P. (2019). On the turing completeness of modern neural network architectures. *ArXiv Preprint ArXiv:1901.03429*.
Peters, M. A., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife, 4*, Article e09651.
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.
Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences, 50*(1), 132–150.
Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology, 16*(10), Article e1008215.
Toker, D., & Sommer, F. T. (2019). Information integration in large brain networks. *PLoS Computational Biology, 15*(2), Article e1006807.
Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition, 79*, Article 102877.
Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences, 19*(12), 757–770.
Usher, M. (2021). Refuting the unfolding-argument on the irrelevance of causal structure to consciousness. *Consciousness and Cognition, 95*, Article 103212.
Watson, J. B., & McDougall, W. (1929). *The battle of behaviorism: An exposition and an exposure*. WW Norton & Company.
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
Wildenhain, T. (2017). On the Turing Completeness of MS PowerPoint. *The Official Proceedings of the Eleventh Annual Intercalary Workshop about Symposium on Robot Dance Party in Celebration of Harry Q Bovik's, 2*, 102–106.
Zeki, S. (2007). A theory of micro-consciousness. *The Blackwell Companion to Consciousness*, 580–588.