


Special Issue: Consciousness science and its theories

Formalizing falsification for theories of consciousness across computational hierarchies

Jake R. Hanson,^{1,2} and Sara I. Walker ^{1,2,3,4,*}

¹School of Earth and Space Exploration, Arizona State University, 550 East Tyler Mall, Tempe, AZ 85287, USA;

²BEYOND Center for Fundamental Concepts in Science, Arizona State University, P.O. Box 870506, Tempe, AZ 85287, USA; ³ASU-SFI Center for Biosocial Complex Systems, Arizona State University, 1031 S. Palm Walk Tempe, AZ 85281-2701, USA; ⁴Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA

*Correspondence address. BEYOND Center for Fundamental Concepts in Science Arizona State University P.O. Box 870506 Tempe, AZ 85287-0506, USA. Tel: (480) 965 - 3240; E-mail: sara.i.walker@asu.edu

Abstract

The scientific study of consciousness is currently undergoing a critical transition in the form of a rapidly evolving scientific debate regarding whether or not currently proposed theories can be assessed for their scientific validity. At the forefront of this debate is Integrated Information Theory (IIT), widely regarded as the preeminent theory of consciousness because it quantified subjective experience in a scalar mathematical measure called Φ that is in principle measurable. Epistemological issues in the form of the “unfolding argument” have provided a concrete refutation of IIT by demonstrating how it permits functionally identical systems to have differences in their predicted consciousness. The implication is that IIT and any other proposed theory based on a physical system’s causal structure may already be falsified even in the absence of experimental refutation. However, so far many of these arguments surrounding the epistemological foundations of falsification arguments, such as the unfolding argument, are too abstract to determine the full scope of their implications. Here, we make these abstract arguments concrete, by providing a simple example of functionally equivalent machines realizable with table-top electronics that take the form of isomorphic digital circuits with and without feedback. This allows us to explicitly demonstrate the different levels of abstraction at which a theory of consciousness can be assessed. Within this computational hierarchy, we show how IIT is simultaneously falsified at the finite-state automaton level and unfalsifiable at the combinatorial-state automaton level. We use this example to illustrate a more general set of falsification criteria for theories of consciousness: to avoid being already falsified, or conversely unfalsifiable, scientific theories of consciousness must be invariant with respect to changes that leave the inference procedure fixed at a particular level in a computational hierarchy.

Keywords: theories and models; computational modeling; consciousness; automata theory; unfolding; integrated information theory

Introduction

Whether or not theories for consciousness can be brought within the purview of science is a subject of intense debate and equally intense importance. The resolution of this debate is necessary for validating theory against experiments in human

subjects. It is also critical to recognizing and/or engineering consciousness in nonhuman systems such as machines. Currently, there is a global, multi-million dollar effort devoted to scientifically validating or refuting the most promising candidate theories, specifically Integrated Information Theory (IIT) and the Global Neuronal Workspace Theory (Reardon 2019). At the same

Received: 17 October 2020; Revised: 22 March 2021. Accepted: 7 April 2021

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

time, it is becoming increasingly unclear whether these theories meet the required scientific criteria for validating them.

Since the early 1990s, scientific studies of consciousness have primarily focused on identifying spatiotemporal patterns in the brain that correlate with what we intuitively consider to be a conscious experience. This is due in large part to advances in medical imaging such as electroencephalograms and functional magnetic resonance imaging (fMRI) that assess brain activity during different functional behaviors (e.g. sleeping, verbal reports, and so on). The empirical data that results from such tests provide evidence for links between spatiotemporal patterns and inferred conscious states. These links, known as Neural Correlates of Consciousness (NCCs), are well-established and form the basis for an entire subfield of contemporary neuroscience (Metzinger 2000; Rees et al. 2002). Despite the success of NCCs, however, there is an underlying epistemic issue with the scientific study of consciousness because conscious states are never directly observed within the NCC framework. Instead, they must be *inferred* based on our own phenomenological experience. For example, when a person is asleep we infer they are less conscious than when they are awake because we have the first-hand subjective experience of what it is like to be both asleep and awake.

While the epistemic issues associated with NCCs are widely known and discussed, the debate around the possibility of falsifying some of the leading theories of consciousness has recently intensified. This resurgence of interest in what constitutes a valid theory for consciousness is primarily due to the new formalization of the scientific issues in the form of “unfolding” arguments (Doerig et al. 2019; Hanson and Walker 2019; Kleiner and Hoel 2021). In particular, the original unfolding argument as clarified by Doerig et al. points to deep logical problems with any causal structure theory (CST) that assumes consciousness supervenes on a particular causal structure independent of outward functional consequences (Doerig et al. 2019), which implies NCCs would be inadequate to validate such theories. Because the currently leading candidate theory for consciousness, IIT, is itself a CST, this has major implications for how we approach the problem of consciousness. To understand how the unfolding argument aims to falsify IIT, it is important to first understand how IIT is constructed as a theory that is derived from simple axioms that make assumptions regarding what conscious experience is, and from these derives a mathematical measure of integrated information Φ that is proposed as a quantification of consciousness. Among the axioms of the theory is the integration axiom, which states that we experience consciousness as an “undivided whole,” meaning, e.g. that our left and right visual fields are integrated into a single conscious experience. Crucially, integration (and the other phenomenological axioms of IIT) must have a direct translation in terms of mathematical machinery to construct the formal theory. For integration, this is achieved by enforcing integration of the physical substrate(s) that gives rise to consciousness, where the precise mathematical definition is in terms of the presence of feedback between the physical components in a system (e.g. neurons). Consequently, any system that is strictly feed-forward is unconscious, by definition in IIT, due to an assumed inability for such physical structures to generate a unified subjective experience. What the unfolding argument showed was that the input-output behavior of any conscious system with feedback and $\Phi > 0$ can be perfectly emulated by a strictly feed-forward system with $\Phi = 0$. To do so, one simply needs to “unfold” the feedback present in the causal structure of the conscious system in a way that preserves the underlying

functionality of the system (i.e. the input-output behavior)—a feat that can be accomplished in the forward or backward direction using feed-forward and recurrent neural networks, respectively (Doerig et al. 2019) or Krohn-Rhodes decomposition (Hanson and Walker 2019). The unfolding argument highlights a key issue with IIT and other potential CSTs: the physical process that is assumed to be causally responsible for generating consciousness does not necessarily correlate with any particular input-output behavior, meaning it is not possible to directly test predictions from the theory.

The scope of this argument is not strictly limited to CSTs. Kleiner and Hoel recently proved that any candidate theory that treats inference and prediction procedures independently must ultimately be subject to the same consequences as theories that succumb to the unfolding argument (Kleiner and Hoel 2020). The validity of their proof rests on their definition of independence, which states that inference and prediction are independent if and only if one can fix the results from the inference procedure while simultaneously allowing predictions from the theory to vary. This results in a *a priori* or “pre-falsification” of the theory, as the theoretical existence of different predictions under fixed inference content necessarily implies that at least one of the predictions is misaligned with the results from the inference procedure. Thus, the dilemma is how one can create a theory of consciousness that simultaneously does not vary under fixed inference content and also does not explicitly depend on the inference content. To satisfy this dual requirement, a viable theory of consciousness must make predictions based on a part of the data set that is kept separate from that which is used to draw inferences (e.g. inferences based on verbal report while predictions are based on fMRI data). Yet, predictions must also match the results from the inference procedure if the theory is to avoid empirical falsification. If inferences are based on input-output behavior, as is commonly assumed, this implies predictions from the theory must be invariant with respect to any change that leaves the input-output behavior of the system fixed. This requirement is quite general and applies beyond CSTs.

The arguments by Doerig et al. and Kleiner and Hoel have addressed the epistemic issues surrounding falsification of theories of consciousness in the abstract. Here, we seek to ground these abstract arguments in a concrete, easily visualizable system that allows clear demonstration of their consequences. The key contribution of the current work is to demonstrate how the issue of falsification is related to the level in the computational hierarchy at which one assesses the validity of a theory for consciousness. To do so, we introduce a hierarchy of formal descriptions that can be used to describe a given finite-state machine. We show that the discrepancy between whether IIT is falsified or unfalsifiable ultimately depends on the computational scale at which inference of subjective experience is made. In particular, we construct isomorphic causal structures (digital circuits) designed to operate a simple electronic tollbooth with and without feedback. In light of this isomorphism, we evaluate the falsification of IIT at two levels of computation for this circuit: at the finite-state automaton (FSA) level and the combinatorial-state automaton (CSA) level and show how the theory is either unfalsifiable at the CSA level or pre-falsified at the FSA level. Our case study demonstrates how candidate measures of consciousness must be invariant with respect to changes in formal descriptions that exist below the level of the specified inference procedure if they are to avoid *a priori* falsification. An added consequence is that our approach provides a window into a deep connection between the current debate surrounding

formalization of falsification arguments and the foundations of computer science. We conclude with a brief discussion regarding what a candidate measure of consciousness that satisfies this constraint might look like, as well as the scope of its applicability.

Results

Defining falsification for theories of consciousness

Falsification is formally defined as a mismatch between a theoretical prediction and an observation. It is an essential component for a theory to be considered scientific in a Popperian framework (Popper 2014). Falsification for theories of consciousness, however, is immediately problematic due to an inability to observe conscious states directly. Instead, they must be *inferred* based on other empirical observations. Thus, falsification for theories of consciousness is defined as a mismatch between prediction and *inference* rather than prediction and observation (Kleiner and Hoel 2020). Consequently, it is possible to disagree as to whether or not a theory of consciousness is falsified due to discrepancies between inference procedures being applied to empirical observations (i.e. the empirical data are the same but inferences are different), or worse, to select inference procedures in accordance with predictions from the theory. This issue highlights the main flaw with the Popperian notion of falsification; namely, there is no such thing as a theory-agnostic inference procedure. Inference based on input-output behavior, e.g. is itself premised on a theory of consciousness grounded in assumptions about human consciousness under normal circumstances, or the belief that “consciousness is as consciousness does” (Turing 1950; Harnad 1991). Indeed, even Popper conceded that logic can never force a scientist to give up a particular theory in the face of surprising observations, meaning falsification can never be proven (Godfrey-Smith 2009). One can always reject the assumptions underlying the inference procedure rather than those underlying the prediction.

Consensus agreement on falsification can only be achieved with respect to a given inference procedure. For example, if a physical system can be transformed into another physical system in a way that preserves the results from an agreed-upon inference procedure while changing the underlying prediction from the theory then a theory of consciousness is falsified with respect to that inference procedure, as this guarantees a mismatch between prediction and inference for at least one of the physical systems under consideration (Kleiner and Hoel 2020). Crucially, it is not necessary to conduct a laboratory experiment in order to falsify a theory, as it is possible to demonstrate the existence of transformations that change the prediction from the theory under fixed inference content without the need to realize these transformations in practice. In such cases, the theory is falsified *a priori* or “pre-falsified” in the language of Kleiner and Hoel 2020. It is pre-falsification that is exploited by Doerig et al. (2019) in the unfolding argument: the input-output behavior of a system is fixed and the underlying causal structure is transformed in a way that changes the predicted Φ value from IIT. If one assumes that the inference procedure takes place at the level of input-output behavior, then the preservation of the input-output behavior fixes the inferred conscious experience and falsifies any and all theories of consciousness that are not invariant under this transformation.

The computational hierarchy

It is typically assumed that inference of conscious states takes place at the level of input-output behavior, as this is the level of description where observations can be intuitively compared to one’s own phenomenal experience. Indeed, the entire field of NCGs rests on the assumed connection between specific output behavior (sleep, self-report, and so on) and inferred conscious states. However, this is not the only formal level of description at which inferences can be made, nor is it immediately apparent that it is optimal. It is at least plausible that lower level descriptions, ranging from neuronal firing to thermodynamic efficiency, could play a role in a valid scientific theory of consciousness. For this reason, our formalism is agnostic to the specific level at which inferences are made; instead, we focus on explicitly characterizing the spectrum of possibilities and examining each in turn. To do this, we introduce the following formal hierarchy that can be used to describe a given computational system, allowing us to precisely formalize the computational level at which a theory is making inferences and predictions.

At the top of the hierarchy is the functional relationship between the inputs, outputs, and internal states that define a computation. These states are typically described in terms of input-output behaviors (“stop,” “walk,” “go,” and so on) but what really gives them meaning mathematically is only their topological relationship with one another. This implies that at this level, the formal description of the computation is not grounded in any particular representation and could, in fact, be realized by radically different logical architectures (Fig. 1). This abstract definition of a computational system corresponds to what Chalmers refers to as the “FSA” level of description, due to the fact it is defined in terms of a global FSA (Chalmers 1993). To add a layer of realizability to this description, one must use what Chalmers refers to as a CSA. In a CSA description, the abstract FSA description of a system is prescribed a specific labeling scheme or encoding of the subcomponents that comprise the global system. In digital electronics, as well as models of the human brain, this encoding is usually given in terms of binary labels that are used to represent abstract functional states in the system. Consequently, transitions between states in the CSA description fix local dependencies between elements, as the correct Boolean function must be applied to each “bit” or “neuron” based on the global state of the system. In addition, the CSA description provides a minimum bound on the memory resources required to run the computation, as the binary labels specify the number of bits required to instantiate the computation. The final level of the computational hierarchy specifies the logic by which the CSA description is implemented. In a Boolean system, this amounts to a specific choice of logic gates used to realize each Boolean function. It is this final level in the hierarchy that we deem the “causal structure” of the system as it fully constrains the logical mechanisms that realize the desired computation (e.g. it describes a digital circuit). Just as there are many CSA representations for a given FSA, there are many “causal structures” (choice of logic functions) for a given CSA. For example, the same CSA description can always be realized using AND, OR, and, NOT gates or universal NAND gates as these examples both form a complete logical basis for Boolean computation. This choice of logic has meaningful consequences in terms of the minimum work required to implement the computation, regardless of the exact physical substrate (Wolpert and Kolchinsky 2020), which may be relevant for theories of consciousness founded on

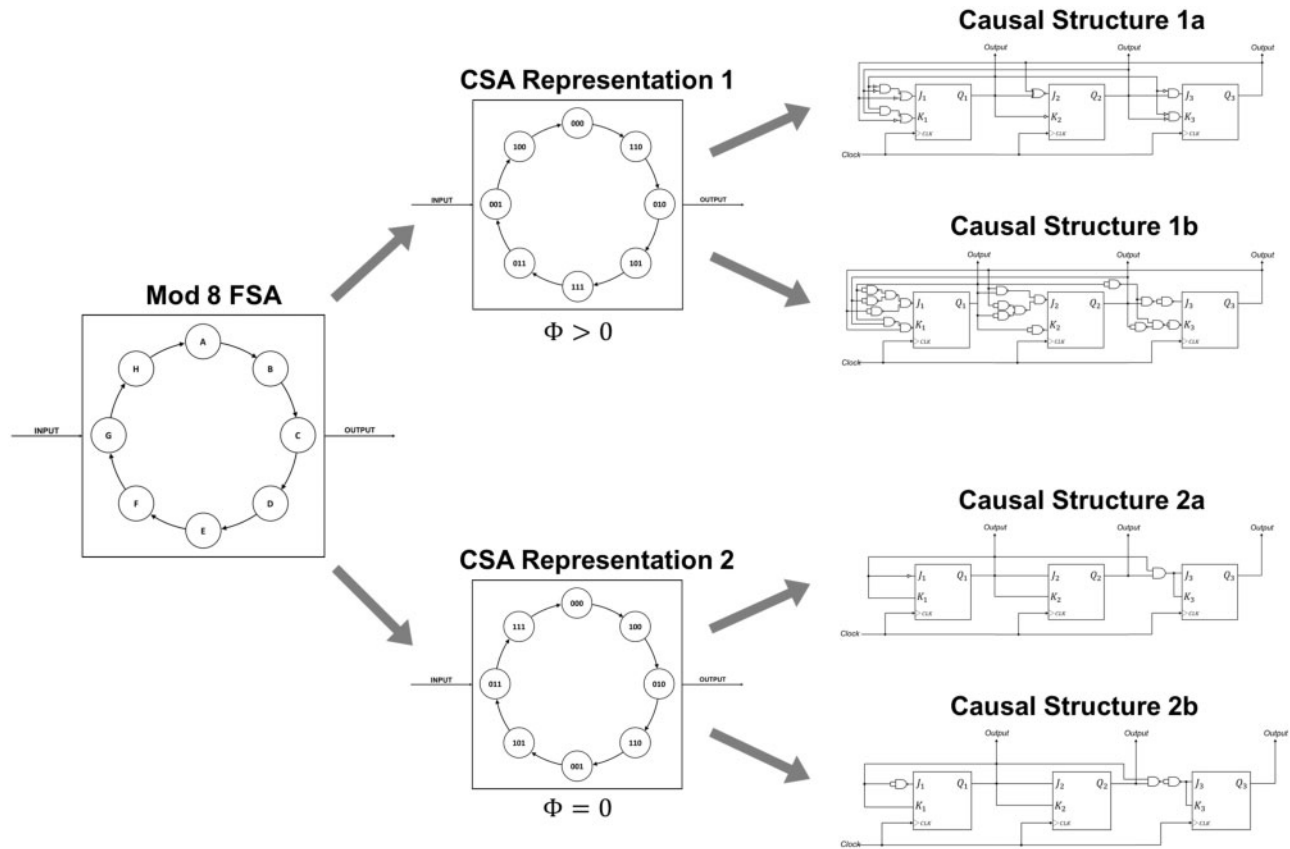


Figure 1. The computational hierarchy used to classify formal levels of description. At the top of the hierarchy is the abstract FSA description of a computation which, in this case, is a mod-eight counter. Beneath this level is the CSA description in which abstract states of the FSA have been assigned specific binary labels which, in turn, constrain local dependencies between subcomponents. Note, it is this level of the hierarchy that IIT uses to calculate Φ . At the bottom of the hierarchy is the full causal structure, as specified in terms of the specific logic gates that implement the Boolean functions from the CSA level. In this case, we have shown two different choices for a complete logical basis: AND/OR/NOT gates or universal NAND gates.

evolutionary arguments of efficiency. However, it is important to note that biological systems typically operate many orders of magnitude above this theoretical limit on thermodynamic efficiency due to the macroscopic size of the components implementing the computation (Bennett 1982). Thus, it is possible to extend our computational hierarchy to include such properties (i.e. the material properties of a given causal structure), but our focus here is primarily on mathematical theories of consciousness for which an abstract description of the system must suffice.

Also of note is the relationship between the computational hierarchy and the ideas of coarse-graining and black-boxing (Hoel 2017; Marshall 2018). Put simply, both coarse-graining and black-boxing refer to the process of throwing away microscopic information in favor of a simplified macroscopic description. The primary difference between coarse-graining and black-boxing is that the former results in a mathematical partition of the microscopic state space (i.e. each microstate is assigned to a macrostate), whereas the latter allows for exogenous factors and the dropping of microstates from the macroscopic description (Hoel 2017). Technically, a partition does not require dimensionality reduction, meaning a one-to-one map (isomorphism) between micro and macro is a valid coarse-graining (We are only talking about isomorphisms between mathematical structures here and are not considering constraints that may

arise as a consequence of physical structure). Thus, the two CSA representations shown in Fig. 1 are technically coarse-grainings of one another, and of the FSA description, as all are in one-to-one correspondence. However, this notion of “coarse-graining” violates our colloquial understanding, especially prevalent in physics, of the term as a process that throws away information in favor of a higher level description. A more appropriate use of the idea is the consideration of a function that partitions the state space of all possible CSAs into equivalence classes based on the FSA that they realize. This results in a many-to-one map from the microscopic state space of all possible CSAs to the macroscopic state space of FSAs, where microstates within a given macrostate share a meaningful similarity in that they are lower level implementations of the same FSA. Indeed, this notion of coarse-graining can be applied equally well to subsequent levels in the computational hierarchy; e.g. the state space of all possible digital circuits (causal structures) can be partitioned into equivalence classes based on the CSA that they implement. Thus, the computational hierarchy can be viewed as a formal coarse-graining of a computation, from its most general representation to its most specific logical implementation. Notably, the number of nodes in the state transition diagram does not change as you move across levels. This is what separates the notion of coarse-graining a computation in what we implement here, from that of coarse-graining in physics.

Prediction and inference within IIT

The primary goal of a mathematical theory of consciousness is to predict whether or not a system is conscious based on a mathematical description of the system. In IIT, the relevant mathematical input for the prediction function is a transition probability matrix (TPM) specifying the conditional probabilities of internal state transitions, and the relevant mathematical output is a scalar value Φ corresponding to the system's overall level of conscious experience. To calculate Φ , the states of the system must be specified as binary strings (Balduzzi and Tononi 2008; Oizumi et al. 2014), which implies the internal representations of functional states must be fixed. In other words, Φ is sensitive to the specific binary labels being used to represent the functional states of an FSA description, which suggests it is the CSA level of the computational hierarchy that is relevant to prediction. Indeed, a CSA is nothing more than a graphical representation of a TPM as it specifies the conditional probabilities of transitions between labeled binary states. In practice, one can also use the causal structure of a system (e.g. a digital circuit) to calculate Φ , but only because the causal structure completely specifies a labeled TPM. Different causal structures with the same CSA description necessarily have the same Φ value, as the TPM is fixed by the CSA representation rather than the causal structure.

Once a CSA is specified, its Φ value and all other relevant predictions from the theory are fixed. However, prior to this, one has a choice of the spatiotemporal scale at which to consider the dynamics. e.g. high-resolution time-series data may be sampled, or spatially coarse-grained data might be instead sampled in order to create a lower dimensional representation of the same system. Interestingly, the Φ value for a given system is not robust to this coarse-graining procedure, meaning the spatiotemporal scale at which one views the dynamics is relevant in predicting whether or not a system is conscious (Hoel et al. 2016). To address this issue, an additional optimization step must be performed wherein all possible spatiotemporal coarse-grainings of a given system (note spatiotemporal coarse-grainings can be different than computational coarse-grainings per discussion above) are considered and the one with the highest Φ value is predicted as the scale in which consciousness resides, in accordance with the postulates of IIT (Hoel et al. 2016).

Unlike prediction, there is no clear prescription in IIT as to where in the computational hierarchy one should *infer* conscious experience, because the theory must be used to predict rather than infer conscious states such that the definition of independence is satisfied. Indeed, confusion over the level at which inference procedures take place plays a prominent role in the on-going debate and confusion surrounding whether or not IIT is an experimentally falsifiable theory. On the one hand, proponents of IIT design experiments to test theoretical predictions against the traditionally held notion that certain outward behaviors such as sleep and self-report are accurate reflections of particular subjective experiences based on our own phenomenal experience. In this case, the inference procedure being used is based on abstract input-output behavior (i.e. the FSA level) where functional states such as sleep are expected in response to inputs such as anesthetics (Casali et al. 2013; Reardon 2019). Crucially, these functional states used for inference do not have natural binary representations and, therefore, can be internally encoded in a variety of different ways with a variety of different causal structures. Thus, inferences are made independently of both the CSA and causal structure descriptions in these experiments. On the other hand, proponents of IIT claim

that it is possible to fix the input-output behavior of a system while still inferring a difference in subjective experiences (e.g. justifying the existence of “philosophical zombies”) (Oizumi et al. 2014; Albantakis and Tononi 2019). In this case, it is the CSA rather than the FSA level of description that must be used to infer the conscious state of a system, as fixed input-output behavior implies a fixed FSA description. Thus, the inference procedure that is used to support the experimental validity of IIT in a traditional laboratory setting must ultimately be rejected in defense of philosophical zombies, which is the paradox on which the unfolding argument is founded (Doerig et al. 2019).

A concrete example

We now turn to a concrete example that demonstrates the unfolding argument as it applies to IIT, and the more general problem of separating prediction from inference, using readily available tabletop electronics. In particular, we will construct isomorphic digital circuits with and without feedback designed to operate a simple electronic counter, such as the tollbooth shown in Fig. 2. See [Supplementary Materials](#) for a descriptions of the methods. Focusing on feedback, as opposed to some other difference in causal structure, allows us to ground our thinking in the specifics of IIT, though the implications of our results readily generalize to any mathematical theory of consciousness (Rescorla 2020).

The FSA description of the tollbooth's behavior is defined by the requirement that it must lift the boom barrier in response to the receipt of exactly eight quarters, as shown schematically in Fig. 2a. To do this, the circuits governing the behavior of the tollbooth must transition through eight internal memory states, corresponding to the eight functional states in the FSA description of the machine, as shown in Fig. 2b. At the CSA level, we insist that both the circuit with feedback and the circuit without feedback be constructed on a three-bit logical architecture, which serves to enforce a strict isomorphism (one-to-one map) between internal states in the two different descriptions. Thus, the FSA description of the two circuits is identical, while the CSA descriptions preserve the topological relationship between inputs, outputs, and internal states. Insisting on isomorphic rather than homomorphic representations provides the tightest possible control on confounding factors that could be used to justify a difference in subjective experience, such as memory allocation (i.e. the number of bits required to instantiate the computation) (Oizumi et al. 2014).

In what follows, we first construct a “conscious” circuit with feedback (and $\Phi > 0$), followed by a functionally identical but “unconscious” circuit with strictly feed-forward connections (and $\Phi = 0$). The general construction of both circuits is the same: first, we assign binary labels to the functional states of the system; then, we map these binary state transitions onto JK flip-flops, which are the bits of our digital circuits; and last, we use Karnaugh Maps to simplify the logic tables of the JK flip-flops in a way that results in simple elementary logic gate operations (e.g. AND, OR, and XOR). As we show, the presence or absence of feedback in the system ultimately stems from the initial choice of the binary labels used to *represent* or *encode* the eight functional states of the system, in accordance with the claim that Φ acts at the CSA level of description. For the system with feedback, we randomly assign these labels in a way that happens to result in $\Phi > 0$ for all states. For the feed-forward system, however, we carefully decompose the underlying functional topology in a way that exploits hierarchical

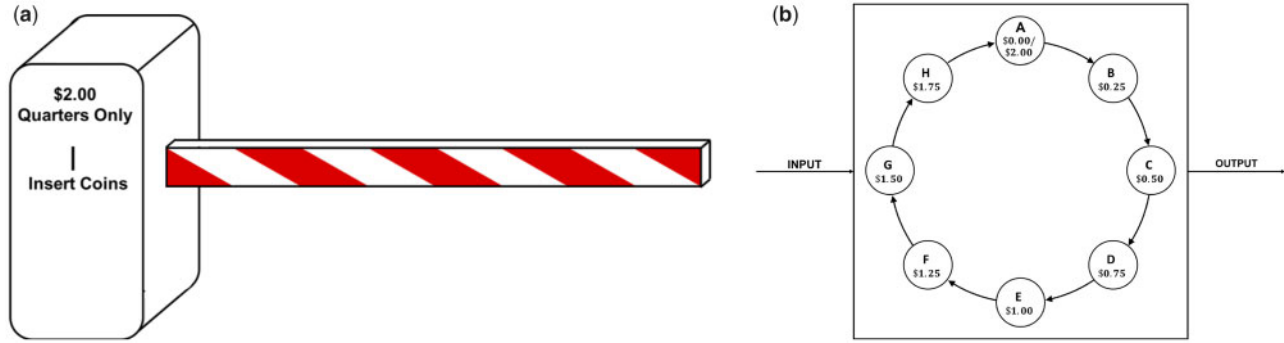


Figure 2. Schematic illustration of a simplified electronic tollbooth (a) and its FSA description (b). The general behavior of the tollbooth is to lift a boom barrier upon receipt of eight quarters (\$2.00). To do this requires the ability to cycle through eight internal memory states $\{A, B, \dots, H\}$, sending each internal state as output to the boom barrier.

relationships such that information flows strictly unidirectionally between components in the system and Φ is guaranteed to be zero (Hanson and Walker 2019). Note, for the tollbooth to function correctly, the boom barrier must be programmed to recognize the internal state A as functionally important, as this is the output that causes the boom barrier to lift and reset. To avoid confusion over this issue, we simply fix the binary representation of state A as 000 across CSA representations, corresponding to the notion that the motor hardware of the boom barrier is programmed to recognize this specific signal as meaningful. In reality, it is typically assumed that the motor hardware can be reprogrammed to recognize any signal as “meaningful”: all that is relevant from a functional perspective is consistency between a circuit and its motor hardware.

Constructing a “conscious” tollbooth

To construct the conscious tollbooth, we randomly assign the following binary labels to represent the eight functional states in the FSA description of the tollbooth:

$$A = 000, B = 110, C = 010, D = 101, E = 111, F = 011, G = 001, H = 100$$

This assignment of labels fully specifies the CSA description of the system, as each binary component (bit) now must transition in accordance with the global state of the system. For example, the transition from state A to state B now requires that the first component of the system transitions from binary state 0 to binary state 1 when the system is in global state 000, which is a constraint on the causal structure. Similarly, the transition from state B to state C specifies that the first component of the system must transition from 1 to 0 when the system is in global state 110. Taken together, the constraints on each individual component in the system at each moment in time generate a truth table that specifies the interdependence between elements and, consequently, the Φ value.

To construct the causal architecture, we must specify the elementary building blocks of our system. In a human brain, these building blocks would be neurons but in a digital circuit, these building blocks are “JK flip-flops,” which are binary memory storage devices (bits) widely used in the construction of basic digital counters (Moore 1958; Cavanagh 2018). The behavior of a JK flip-flop is quite simple: there are two stable internal memory states (0 and 1), two input channels (the J input and the K input), and a “clock” that serves to synchronize multiple flip-flops within a circuit. Upon receipt of voltage on a line from the clock, the flip-flop does one of four things depending on the

state of the J and K input channel: if the JK input is 00 the internal state remains unchanged (“latch”), if the JK input is 01 the internal state resets to 0 (“reset”), if the JK input is 10 the internal state is set to 1 (“set”), and if the JK input is 11 the internal state is flipped (“toggle”). Thus, for any given internal state transition— $Q_i(t_0) \rightarrow Q_i(t_1)$ —there are two different pairs of JK input that will correctly realize the transition, as shown in Fig. 3. This degeneracy provides flexibility when it comes to the design of the elementary logic gate operations required to realize the underlying Boolean logic.

With the specification of the binary labels and the choice of electronic components, we can now finish the construction of the causal structure in terms of elementary logic gates. To do so, we first convert the state transitions of each individual component into their associated JK values. As mentioned, there is degeneracy in the choice of JK input which means we only have to specify one of the input channels (either J or K) to get the desired transition. For each component in the circuit, there is a column in Fig. 4a corresponding to the JK value that is required; note, inputs that do not need to be specified are denoted with an asterisk. Next, we must determine the elementary logic gates required to get the correct JK transitions given the current state of the system. For instance, when the system is in global state 110, the value of K_1 (the K-input to the first component) must be 1, but when the system is in global state 111 the value of K_1 must be 0. Taken together, the eight states of the system comprise a truth table of JK input as a function of the global state of the system, as shown in Fig. 4b. Ordering these truth tables in gray code yields “Karnaugh maps,” which allow straightforward identification of the elementary logic gates required to operate the circuit (Karnaugh 1953). The elementary logic expression for each of the six input channels, in terms of AND, OR, XOR, and

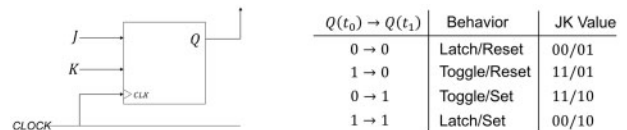


Figure 3. A JK flip-flop is a widely used binary memory device (bit) in digital electronics (a). The internal state of the flip-flop takes one of two values ($Q \in \{0, 1\}$) and is continuously sent as output. Upon receipt of a voltage from a clocked input, the voltages on the two input channels J and K dictate the state transitions of Q (see main). For any desired internal state transition $Q(t_0) \rightarrow Q(t_1)$, there are two JK inputs that will correctly realize the transition (b) which provides flexibility when it comes to circuit design.

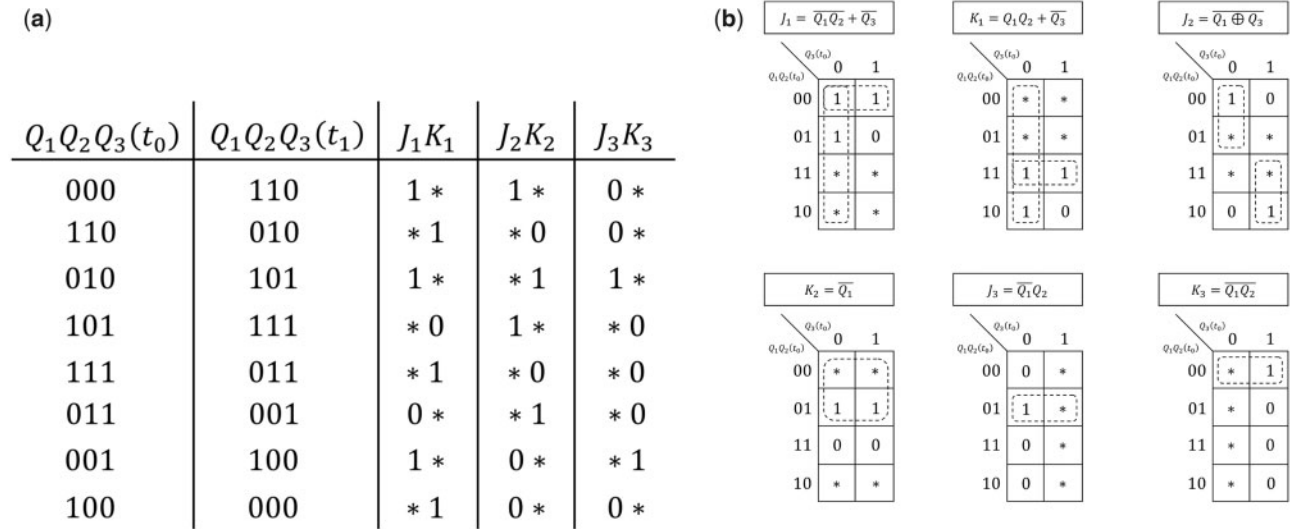


Figure 4. To construct the digital circuit for the conscious tollbooth, we convert the global state transitions into their associated JK values (a). Then, we use Karnaugh maps to determine the elementary logic required to update each component (b). The presence of feedback in the resultant digital circuit is evident by the dependence of earlier components on later components (e.g. $J_1 = \overline{Q_1}Q_2 + \overline{Q_3}$) and vice versa (e.g. $K_3 = \overline{Q_1}Q_2$).

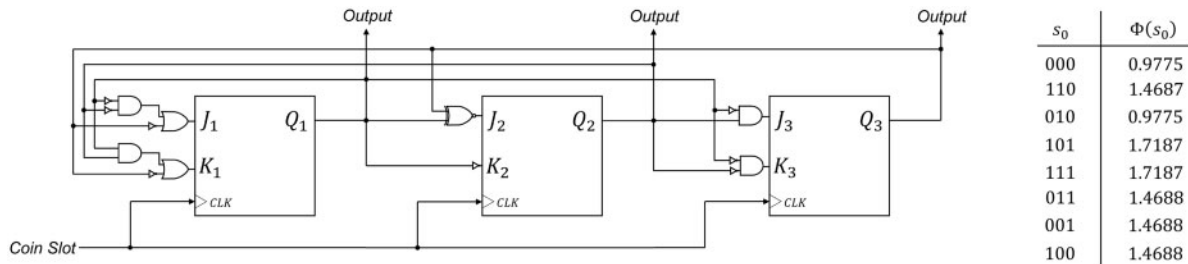


Figure 5. An integrated digital circuit (a) designed to operate the electronic tollbooth was shown in Fig. 2. As can be seen, the causal structure contains meaningful feedback in the form of bidirectional dependencies between pairs of elements and, consequently, has $\Phi > 0$ for all states (b).

NOT gates, is shown above the corresponding Karnaugh map in Fig. 4b.

The elementary logic expressions for the behavior of each JK input complete the construction of our circuit, which is shown in Fig. 5a. Clearly, this circuit contains meaningful feedback between components, as the state of the first component depends on the state of the second and third and vice versa. The last thing to check is whether or not this feedback is associated with the presence of consciousness according to IIT, as feedback is a necessary but not sufficient condition for $\Phi > 0$. Using the python package PyPhi (Mayner et al. 2018), we find $\Phi > 0$ for all states (Fig. 5b), meaning this tollbooth is indeed considered conscious according to IIT.

Constructing an “unconscious” tollbooth

In the previous section, we demonstrated the construction of a causal structure designed to operate the electronic tollbooth shown in Fig. 2a. We did so by randomly assigning 3-bit binary labels to represent the function states ($\{A, B, \dots, H\}$) of the system and constructing the logic of the digital circuit in a way that correctly realizes these labeled state transitions. The result was a circuit that relied on feedback connections (i.e. bi-directional information exchange between components) and had $\Phi > 0$ for all states (Fig. 5). In this section, we demonstrate that it is possible to assign binary labels to functional states in a different way, such that the causal structure that results instantiates the

same FSA (Fig. 2b) but does not make use of feedback connections. To do so, we will “unfold” the underlying dynamics of the system in a way that guarantees a causal architecture with $\Phi = 0$ for all states in the system.

The process of unfolding a finite-state description of a system is based on techniques closely related to the Krohn-Rhodes theorem from automata theory, which states: any abstract deterministic finite-state automata (FSA) can be realized using a strictly feed-forward causal architecture comprised solely of simple elementary components (Krohn and Rhodes 1965; Zeiger 1967). To do so isomorphically, one must find a “nested sequence of preserved partitions,” which creates a hierarchical labeling scheme wherein earlier components (flip-flops) transition independently of later components (Zeiger 1968; Hanson and Walker 2019). Due to this hierarchical independence, information is guaranteed to flow unidirectionally from earlier components to later components, thereby ensuring a strictly feed-forward logical architecture and, correspondingly, $\Phi = 0$ for all states. While a full discussion of Krohn-Rhodes decomposition is beyond the scope of this study (Egri-Nagy and Nehaniv 2015), we briefly describe the relevant methodology for constructing a nested sequence of preserved partitions in the Methods section. The result, applied to the finite-state description of the tollbooth shown in Fig. 2b, is the following set of binary labels used to encode the functional states of our system:

A = 000, B = 100, C = 010, D = 110, E = 001, F = 101, G = 011, H = 111

Notice, in this labeling scheme, the value of the first component (or “coordinate”) partitions the underlying state space of the system into two macrostates: {A, C, E, G} and {B, D, F, H} and can be thought of as high-level representation of “even” and “odd” states. These macrostates are useful due to the fact they transition deterministically back and forth between one another. Thus, knowing the future state of the first component depends solely on knowing the current state of the first component. Similarly, the future state of the second component is completely deterministic given the current state of the first and second components and is agnostic to the third. In this way, each additional component offers a refined estimate as to where in the global state space the current microstate is located (DeDeo 2011), hence the claim that the labeling scheme is “hierarchical.”

With hierarchical coordinates assigned, the circuit construction now proceeds in a fashion identical to the previous section. Namely, we convert the binary state transitions into their associated JK values, shown in Fig. 6a. Then, we construct truth tables for the state of each J and K input given the global state of the system; and last, we order these truth tables in gray code (Karnaugh Maps) and assign elementary logic gates to each input channel (Fig. 6b). The resulting logical architecture is shown in Fig. 7a. As required, the circuit is strictly feed-forward, as evident by the fact that each component depends solely on itself or earlier components. This, in turn, guarantees $\Phi = 0$ for all states of the system (Fig. 7b) as the presence of feedback connections is assumed to be a necessary condition for consciousness according to IIT.

Discussion

Falsification, unfalsifiability, and the scientific verification of explanations for consciousness

In light of our results, it is clear that IIT predicts a difference in subjective experience between the “conscious” tollbooth in Fig. 5 and the “unconscious” tollbooth in Fig. 7, based on their

difference in Φ value. Thus, falsification is a matter of whether or not one can infer a corresponding difference that justifies this difference in prediction. Because the two systems have the same FSA description, any inference procedure that takes place at the FSA level or above automatically falsifies the theory, as the inference content is fixed at this level [implying a mismatch between prediction and inference for at least one of the two predictions (Kleiner and Hoel 2020)]. Consequently, IIT is falsified with respect to inference procedures that are based on the input–output behavior of the system, as this is the FSA level of description.

This implies that if IIT is to be considered falsifiable (and not already falsified), inference must take place at the CSA level or below. At the CSA level, however, the full utility of the isomorphism is evident as the only mathematical difference between the CSA description with and without $\Phi > 0$ is a permutation of the binary labels used to internally label functional states. This means that ultimately this singular difference must be used for both inference and prediction, which implies prediction and inference must be coupled together in a way that violates independence—rendering the theory unfalsifiable. For example, in Oizumi et al. (2014), the authors argue that the reason functionally indistinguishable systems with different Φ values have justifiably different subjective experiences is because there are different amounts of feedback present in the internal dynamics. Thus, the integration postulate is used to justify the difference in Φ values, neglecting the fact that Φ is derived to be in correspondence with the integration postulate; in other words, the presence or absence of feedback must be used as both a means (prediction) and an end (inference) to avoid falsification.

The inability to falsify a theory does not necessarily imply it is unscientific. Indeed, when considered as a “phenomenology-first” approach, IIT can be argued to be well-grounded rather than circular. In this case, the scientific merit of the theory is in the translation of its phenomenological axioms into empirical predictions (i.e. Φ values) (Negro 2020). Even if these predictions cannot be independently verified, the theory may be considered scientific if it connects with other ideas and is embedded in a larger conceptual structure in a way that exposes the axioms to

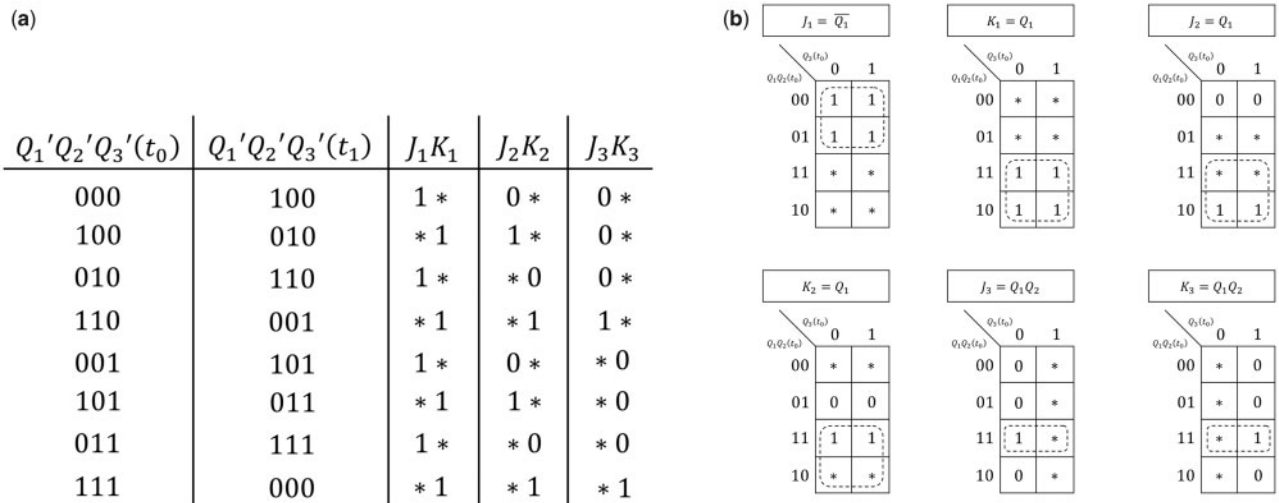


Figure 6. The state transitions and JK values (a) corresponding to the hierarchical labeling scheme described in the main text. Panel (b) shows the Karnaugh maps used to determine the elementary logic gates used in the construction of the feed-forward logical architecture. Note, the logical dependence between components is strictly unidirectional (e.g. J_2 and K_2 depend only on the state of Q_1).

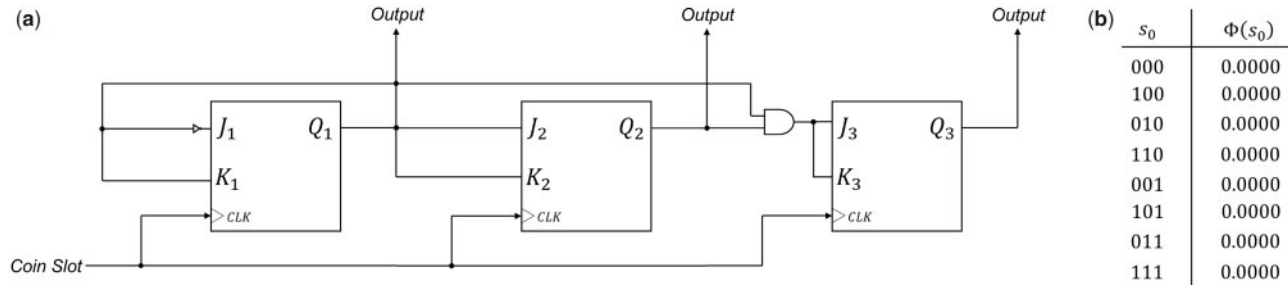


Figure 7. A feed-forward digital circuit (a) designed to operate the electronic tollbooth was shown in Fig. 2. This causal structure operates under the same memory constraints as the integrated circuit (i.e. a three-bit logical architecture) but has $\Phi = 0$ for all states (b).

observation. This approach to the demarcation problem (i.e. what constitutes as science) does not require falsification, as there are many ways in which observations can be used to modify and assess the core principles of a theory without directly falsifying them. However, if the theory is constructed in such a way that it is protected from all risk, then this is unscientific handling of the core principles of the theory (Godfrey-Smith 2009). To scientifically handle the basic principles of IIT is to work out what difference it makes to things we can observe if the principles of IIT are true. What our results show is that the presence or absence of Φ makes no difference to things we can observe, beyond the syntactical differences between internal representations.

Going forward

Our results prove an *a priori* or “pre-falsification” of IIT with respect to inference at the FSA level using a simple, readily realizable model. We have shown that what Φ measures are a consequence of a particular CSA representation (or encoding) of a computation, without clear grounding in terms of phenomenology beyond what is assumed by the postulates of the theory. For a theory to avoid the epistemic problems revealed by IIT under the isomorphic transformation we introduce requires that no transformation or “substitution” exists that changes the prediction without affecting the inference, in complete agreement with previous work related to the unfolding argument (Doerig et al. 2019; Kleiner and Hoel 2020). This, in turn, implies that beneath a specified level of inference, a mathematical theory of consciousness must be invariant with respect to any and all changes that leave the results from the inference procedure fixed, corresponding to the definition of independence from Kleiner and Hoel (2020). Put simply, if you can make a change to a system that does not affect what will be used to infer conscious states, then such a change must not affect the prediction from the theory.

An example of a candidate measure that satisfies these requirements is Group Complexity (Rhodes and Nehaniv 2009). Like Φ , Group Complexity is a measure of computational complexity based on a topological description of a computation. Specifically, it counts the number of resets necessary to complete a Krohn-Rhodes decomposition of the computation (Zeiger 1967; Egri-Nagy and Nehaniv 2008), meaning all integration is decomposed into feed-forward representations prior to the complexity being measured. This, in turn, puts all CSA representations on an equal playing field, as complexity comes in two forms: “resets” and feedback connections. By first unfolding the dynamics of an integrated circuit, one can measure the complexity of the underlying computation at the level of the FSA rather than any particular CSA representation. Consequently, it is invariant with respect to changes below the FSA level, as desired. In light of this, it is

important to ask whether there is anything to be gained from a candidate measure of consciousness such as Group Complexity. In answer, one must first ask whether or not the measure is falsifiable by examining whether inference and prediction can be kept independent. This is easy enough to check for Group Complexity, as inferences are canonically made based on input-output behavior while Group Complexity is a topological measure. Given that there is no *a priori* dependence between a topological description of a computation and the input-output behavior it realizes, GC is indeed capable of producing nontrivial predictions. In regard to whether or not these predictions are falsifiable, it is certainly possible that we infer a conscious state based on input-output behavior that is in disagreement with a prediction from a measure of complexity such as Group Complexity. For example, if the Group Complexity of a model system increases when the system goes asleep, then this serves as falsification with respect to the canonical inference that sleep corresponds to lower subjective experience. While this may sound virtually identical to experiments designed to test IIT (Casali et al. 2013; Reardon 2019), the crucial difference is that Group Complexity is invariant with respect to changes below the FSA level.

Group Complexity is a measure of complexity that is both nontrivial and falsifiable, and therefore, it is an epistemologically sound measure of consciousness that acts on the same mathematical structures and retains some of the original insight that motivated IIT (Tononi and Edelman 1998). Yet, at face value, Group Complexity seems much too simple to truly quantify the conscious experience. For one, it coarse-grains all of the richness associated with sensorimotor experience into a scalar value that retains none of the corresponding physical information associated with specific senses, that is it has no implicit explanation for “what it is like” to be something (Nagel 1974). While IIT deals with this problem by equating multi-dimensional vectors with “concepts in qualia space,” such sophistications are even harder to ground experimentally than a scalar measure, as the ability to empirically validate the nuances of a rich phenomenal structure are limited by our ability to empirically infer such structures. Given this, it seems the biggest problem faced by consciousness research going forward is not necessarily the mathematical structures that a theory can predict but the mathematical structures that one can infer, as ultimately predictions from a theory are only as believable as the inferences that ground them. We know based on first-hand phenomenal experience of consciousness that certain behaviors such as sleep and verbal report are likely accurate reflections of consciousness in human beings and it is these behaviors that must be leveraged by the inference procedure. Beyond these few specific examples, however, it is difficult to imagine what else can be used to infer conscious states that

are not also used to make predictions within the theory. In cases where we lose phenomenological grounding, such as artificial intelligence, this issue is especially problematic (Doerig et al. 2020).

While the inability to test what we assume to be consciousness has always plagued the study of consciousness, we hope that formalizing the problem in terms of the level of computational abstraction at which inferences and predictions take place makes it clear that there are mathematical constraints that all theories of consciousness must satisfy if they are to be falsifiable. Namely, the theory must be invariant with respect to changes that leave the results from the inference procedure unaffected [satisfying the definition of independence from Kleiner and Hoel (2020)]. In IIT, the inference procedure being used to justify the experimental validity of the theory is at the level of the input-output behavior of the system, and therefore Φ must be invariant with respect to equivalence classes that share the same FSA description. The fact that it is not either falsifies the theory or renders it unfalsifiable, depending on which level in the computational hierarchy one uses for inference. Our analyses indicate that not only are new theories of consciousness needed, but new frameworks for assessing the validity of these theories are needed as well. The latter, e.g. could be addressed by constructing theories that do not aim to quantify what subjective experience is, but rather the functional consequences of subjective experience in the physical world (Walker and Davies 2017).

Supplementary Data

Supplementary data is available at NCONSC Journal online.

Acknowledgments

The authors thank the Emergence@ASU team for helpful feedback on this work as well as two anonymous reviewers whose comments significantly improved the manuscript.

Funding

This work was in part supported by the John Templeton Foundation and the Foundational Questions in Science Institute.

Data Availability

All relevant data necessary to reproduce the work in this manuscript are included in the main text and supplement.

Conflict of interest statement. None declared.

References

- Albantakis L, Tononi G. Causal composition: structural differences among dynamically equivalent systems. *Entropy* 2019;21:989.
- Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol* 2008;4:e1000091.
- Bennett CH. The thermodynamics of computation—a review. *Int J Theor Phys* 1982;21:905–40.
- Casali AG, Gosseries O, Rosanova M, et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Trans Med* 2013;5:198ra105.
- Cavanagh J. *Sequential Logic: Analysis and Synthesis*. Boca Raton, FL, USA: CRC Press, 2018.
- Chalmers DJ. *A Computational Foundation for the Study of Cognition*. 1993 (Unpublished).
- DeDeo S. Effective theories for circuits and automata. *Chaos* 2011;21:037106.
- Doerig A, Schurger A, Herzog MH. Hard criteria for empirical theories of consciousness. *Cogn Neurosci* 2020;1–22.
- Doerig A, Schurger A, Hess K, et al. The unfolding argument: why IIT and other causal structure theories cannot explain consciousness. *Conscious Cogn* 2019;72:49–59.
- Egri-Nagy A, Nehaniv CL. Computational holonomy decomposition of transformation semigroups. arXiv preprint arXiv:1508.06345. 2015.
- Egri-Nagy A, Nehaniv CL. Hierarchical coordinate systems for understanding complexity and its evolution, with applications to genetic regulatory networks. *Artif Life* 2008;14:299–312.
- Godfrey-Smith P. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago IL, USA: University of Chicago Press, 2009.
- Hanson JR, Walker SI. Integrated information theory and isomorphic feed-forward philosophical zombies. *Entropy* 2019;21:1073.
- Harnad S. Other bodies, other minds: a machine incarnation of an old philosophical problem. *Minds Mach* 1991;43–54.
- Hoel E. When the map is better than the territory. *Entropy* 2017;19:188.
- Hoel EP, Albantakis L, Marshall W, et al. Can the macro beat the micro? Integrated Information across Spatiotemporal Scales. *Neurosci Conscious* 2016;2016:niw012.
- Karnaugh M. The map method for synthesis of combinational logic circuits. *Trans Am Inst Elect Engin Commun Elect* 1953;72:593–99.
- Kleiner J, Hoel E. Falsification and consciousness. *Neuroscience of Consciousness* 2021;2021:niab001.
- Krohn K, Rhodes J. Algebraic theory of machines. I. prime decomposition theorem for finite semigroups and machines. *Trans Am Math Soc* 1965;116:450–64.
- Marshall W, Albantakis L, Tononi G. Black-boxing and cause-effect power. *PLoS Comput Biol* 2018;14:e1006114.
- Mayner WG, Marshall W, Albantakis L, et al. Pyphi: a toolbox for integrated information theory. *PLoS Comput Biol* 2018;14:e1006343.
- Metzinger T. *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge MA, USA: MIT Press, 2000.
- Moore EF. Logical design of digital computers. *J Symb Logic* 1958;23:363–65.
- Nagel T. What is it like to be a bat? *Philos Rev* 1974;83:435–50.
- Negro N. Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenol Cogn Sci* 2020;1:18.
- Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 2014;10:e1003588.
- Popper K. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York, USA: Routledge, 2014.
- Reardon S. Rival theories face off over brain's source of consciousness. *Science* 2019;366:293.
- Rees G, Kreiman G, Koch C. Neural correlates of consciousness in humans. *Nat Rev Neurosci* 2002;3:261–70.
- Rescorla M. The computational theory of mind. In Zalta, EN (ed.), *The Stanford Encyclopedia of Philosophy (Metaphysics Research Lab, Stanford University)* 2020, Spring 2020 edn. <<https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>>.
- Rhodes J, Nehaniv CL. *Applications of Automata Theory and Algebra*. Singapore: World Scientific, 2009.

- Tononi G, Edelman GM. Consciousness and complexity. *Science* 1998;**282**:1846–51.
- Turing A. Computing machinery and intelligence. *Mind* 1950;**59**:433.
- Walker SI, Davies PC. The 'hard problem' of life. In Walker, SI, Davies PCW, Ellis G.F.R. (eds.), *From Matter to Life: Information and Causality*. Cambridge UK: Cambridge University Press, 2017;19–37.
- Wolpert DH, Kolchinsky A. Thermodynamics of Computing with Circuits. *New Journal of Physics* 2020;**22**:063047.
- Zeiger HP. Cascade decomposition using covers. In Arbib, AM (ed.), *Algebraic Theory of Machines, Languages, and Semigroups*, chap. 4. Academic Press, 1968, 55–80.
- Zeiger P. Yet another proof of the cascade decomposition theorem for finite automata. *Theory Comput Syst* 1967;**1**:225–28.