Falsification and consciousness

Johannes Kleiner ⬤ [1] and Erik Hoel ⬤ [2,*]

[1]Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, Germany; [2]Allen Discovery Center, Tufts University, Medford, MA, USA

*Correspondence address. Allen Discovery Center, Tufts University, Medford, MA, USA. E-mail: erik.hoel@tufts.edu

Abstract

The search for a scientific theory of consciousness should result in theories that are falsifiable. However, here we show that falsification is especially problematic for theories of consciousness. We formally describe the standard experimental setup for testing these theories. Based on a theory's application to some physical system, such as the brain, testing requires comparing a theory's predicted experience (given some internal observables of the system like brain imaging data) with an inferred experience (using report or behavior). If there is a mismatch between inference and prediction, a theory is falsified. We show that if inference and prediction are independent, it follows that any minimally informative theory of consciousness is automatically falsified. This is deeply problematic since the field's reliance on report or behavior to infer conscious experiences implies such independence, so this fragility affects many contemporary theories of consciousness. Furthermore, we show that if inference and prediction are strictly dependent, it follows that a theory is unfalsifiable. This affects theories which claim consciousness to be determined by report or behavior. Finally, we explore possible ways out of this dilemma.

Keywords: consciousness; theories and models

Introduction

Successful scientific fields move from exploratory studies and observations to the point where theories are proposed that can offer precise predictions. Within neuroscience, the attempt to understand consciousness has moved out of the exploratory stage and there are now a number of theories of consciousness capable of predictions that have been advanced by various authors (Koch *et al.* 2016).

At this point in the field's development, falsification has become relevant. In general, scientific theories should strive to make testable predictions (Popper 1968). In the search for a scientific theory of consciousness, falsifiability must be considered explicitly as it is commonly assumed that consciousness itself cannot be directly observed, instead it can only be inferred based off of report or behavior.

Contemporary neuroscientific theories of consciousness first began to be proposed in the early 1990s (Crick 1994). Some have been based directly on neurophysiological correlates, such as proposing that consciousness is associated with neurons firing at a particular frequency (Crick and Koch 1990) or activity in some particular area of the brain like the claustrum (Crick and Koch 2005). Other theories have focused more on the dynamics of neural processing, such as the degree of recurrent neural connectivity (Lamme 2006). Others yet have focused on the "global workspace" of the brain, based on how signals are propagated across different brain regions (Baars 1997). Specifically, Global Neuronal Workspace (GNW) theory claims that consciousness is the result of an "avalanche" or "ignition" of widespread neural activity created by an interconnected but dispersed network of neurons with long-range connections (Sergent and Dehaene 2004).

Another avenue of research strives to derive a theory of consciousness from analysis of phenomenal experience. The most promising example thereof is Integrated Information Theory (IIT) (Tononi 2004, 2008; Oizumi *et al.* 2014). Historically, IIT is the first well-formalized theory of consciousness. It was the first (and arguably may still be the lone) theory that makes precise quantitative predictions about both the contents and level of consciousness (Tononi 2004). Specifically, the theory takes

the form of a function, the input of which is data derived from some physical system's internal observables, while the output of this function is predictions about the contents of consciousness (represented mathematically as an element of an experience space) and the level of consciousness (represented by a scalar value Φ).

Both GNW and IIT have gained widespread popularity, sparked a general interest in consciousness, and have led to dozens if not hundreds of new empirical studies (Massimini *et al.* 2005; Del Cul *et al.* 2007; Dehaene and Changeux 2011; Gosseries *et al.* 2014; Wenzel *et al.* 2019). Indeed, there are already significant resources being spent attempting to falsify either GNW or IIT in the form of a global effort pre-registering predictions from the two theories so that testing can be conducted in controlled circumstances by researchers across the world (Ball 2019; Reardon 2019). We therefore often refer to both GNW and IIT as exemplar theories within consciousness research and show how our results apply to both. However, our results and reasoning apply to most contemporary theories, e.g. (Lau and Rosenthal 2011; Chang *et al.* 2019), particularly in their ideal forms. Note that we refer to both "theories" of consciousness and also "models" of consciousness, and use these interchangeably (Seth 2007).

Due to IIT's level of formalization as a theory, it has triggered the most in-depth responses, expansions, and criticisms (Cerullo 2015; Bayne 2018; Mediano *et al.* 2019; Kleiner and Tull 2020) since well-formalized theories are much easier to criticize than nonformalized theories. Recently, one criticism levied against IIT was based on how the theory predicts feedforward neural networks have zero Φ and recurrent neural networks have nonzero Φ. Since a given recurrent neural network can be "unfolded" into a feedforward one while preserving its output function, this has been argued to render IIT outside the realm of science (Doerig *et al.* 2019). Replies have criticized the assumptions which underlie the derivation of this argument (Tsuchiya *et al.* 2019; Kleiner 2020).

Here, we frame and expand concerns around testing and falsification of theories by examining a more general question: what are the conditions under which theories of consciousness (beyond IIT alone) can be falsified? We outline a parsimonious description of theory testing with minimal assumptions based on first principles. In this agnostic setup, falsifying a theory of consciousness is the result of finding a mismatch between the inferred contents of consciousness (usually based on report or behavior) and the contents of consciousness as predicted by the theory (based on the internal observables of the system under question).

This mismatch between prediction and inference is critical for an empirically meaningful scientific agenda, because a theory's prediction of the state and content of consciousness on its own cannot be assessed. For instance, imagine a theory that predicts (based on internal observables like brain dynamics) that a subject is seeing an image of a cat. Without any reference to report or outside information, there can be no falsification of this theory, since it cannot be assessed whether the subject was actually seeing a "dog" rather than "cat." Falsifying a theory of consciousness is based on finding such mismatches between reported experiences and predictions.

In the following work, we formalize this by describing the prototypical experimental setup for testing a theory of consciousness. We come to a surprising conclusion: a widespread experimental assumption implies that most contemporary theories of consciousness are already falsified.

The assumption in question is the *independence* of an experimenter's inferences about consciousness from a theory's predictions. To demonstrate the problems this independence creates for contemporary theories, we introduce a "substitution argument." This argument is based on the fact that many systems are equivalent in their reports (e.g. their outputs are identical for the same inputs), and yet their internal observables may differ greatly. This argument constitutes both a generalization and correction of the "unfolding argument" against IIT presented in Doerig *et al.* (2019). Examples of such substitutions may involve substituting a brain with a Turing machine or a cellular automaton since both types of systems are capable of universal computation (Turing 1937; Wolfram 1984) and hence may emulate the brain's responses, or replacing a deep neural network with a single-layer neural network, since both types of networks can approximate any given function (Hornik *et al.* 1989; Schäfer and Zimmermann 2006).

Crucially, our results do not imply that falsifications are impossible. Rather, they show that the independence assumption implies that whenever there is an experiment where a theory's predictions based on internal observables and a system's reports agree, there exists also an actual physical system that falsifies the theory. One consequence is that the "unfolding argument" concerning IIT (Doerig *et al.* 2019) is merely a small subset of a much larger issue that affects all contemporary theories which seek to make predictions about experience off of internal observables. Our conclusion shows that if independence holds, all such theories come falsified *a priori*. Thus, instead of putting the blame of this problem on individual theories of consciousness, we show that it is due to issues of falsification in the scientific study of consciousness, particularly the field's contemporary usage of report or behavior to infer conscious experiences.

A simple response to avoid this problem is to claim that report and inference are not independent. This is the case, e.g., in behaviorist theories of consciousness, but arguably also in Global Workspace Theory (Baars 2005), the "attention schema" theory of consciousness (Graziano and Webb 2015) or "fame in the brain" (Dennett 1991) proposals. We study this answer in detail and find that making a theory's predictions and an experimenter's inferences *strictly dependent* leads to pathological unfalsifiability.

Our results show that if the independence of prediction and inference holds true, as in contemporary cases where report about experiences is relied upon, it is likely that no current theory of consciousness is correct. Alternatively, if the assumption of independence is rejected, theories rapidly become unfalsifiable. While this dilemma may seem like a highly negative conclusion, we take it to show that our understanding of testing theories of consciousness may need to change to deal with these issues.

Formal Description of Testing Theories

Here, we provide a formal framework for experimentally testing a particular class of theories of consciousness. The class we consider makes *predictions* about the *conscious experience* of *physical systems* based on *observations or measurements*. This class describes many contemporary theories, including leading theories such as IIT (Oizumi *et al.* 2014), GNW Theory (Dehaene and Changeux 2004), Predictive Processing [when applied to account for conscious experience (Hohwy 2012; Hobson *et al.* 2014; Seth 2014; Clark 2019; Dolega and Dewhurst 2020)], or Higher Order Thought Theory (Rosenthal 2002). These theories may be

motivated in different ways, or contain different formal structures, such as e.g., the ones of category theory (Tsuchiya *et al.* 2016). In some cases, contemporary theories in this class may lack the specificity to actually make precise predictions in their current form. Therefore, the formalisms we introduce may sometimes describe a more advanced form of a theory, one that can actually make predictions.

In the following section, we introduce the necessary terms to define how to falsify this class of theories: how the measurement of a physical system's observables results in datasets (Experiments section), how a theory makes use of those datasets to offer predictions about consciousness (Predictions section), how an experimenter makes inferences about a physical system's experiences (Inferences section), and finally how falsification of a theory occurs when there is a mismatch between a theory's prediction and an experimenter's inference (Falsification section). In Summary section, we give a summary of the introduced terms. In subsequent sections, we explore the consequences of this setup, such as how all contemporary theories are already falsified if the data used by inferences and predictions are independent, and also how theories are unfalsifiable if this is changed to a strict form of dependency.

Experiments

All experimental attempts to either falsify or confirm a member of the class of theories we consider begin by examining some particular physical system which has some specific physical configuration, state, or dynamics, p. This physical system is part of a class P of such systems which could have been realized, in principle, in the experiment. For example, in IIT, the class of systems P may be some Markov chains, set of logic gates, or neurons in the brain, and every $p \in P$ denotes that system being in a particular state at some time t. On the other hand, for GNW, P might comprise the set of long-range cortical connections that make up the global workspace of the brain, with p being the activity of that global workspace at that time.

Testing a physical system necessitates experiments or observations. For instance, neuroimaging tools like fMRI or EEG have to be used in order to obtain information about the brain. This information is used to create datasets such as functional networks, wiring diagrams, models, or transition probability matrices. To formalize this process, we denote by $\mathcal{O}$ all possible datasets that can result from observations of P. Each $o \in \mathcal{O}$ is one particular dataset, the result of carrying out some set of measurements on p. We denote the datasets that can result from measurements on p as obs(p). Formally:

$$\text{obs} : P \twoheadrightarrow \mathcal{O}, \tag{1}$$

where obs is a *correspondence*, which is a "generalized function" that allows more than one element in the image obs(p) (functions are a special case of correspondences). A correspondence is necessary because, for a given p, various possible datasets may arise, e.g., due to different measurement techniques such as fMRI vs. EEG, or due to the stochastic behavior of the system, or due to varying experimental parameters. In the real world, data obtained from experiments may be incomplete or noisy, or neuroscientific findings difficult to reproduce (Gilmore *et al.* 2017). Thus for every $p \in P$, there is a whole class of datasets which can result from the experiment.

Note that obs describes the experiment, the choice of observables, and all conditions during an experiment that generates the dataset o necessary to apply the theory, which may

differ from theory to theory, such as interventions in the case of IIT. In all realistic cases, the correspondence obs is likely quite complicated since it describes the whole experimental setup. For our argument, it simply suffices that this mapping exists, even if it is not known in detail.

It is also worth noting here that all leading neuroscientific theories of consciousness, from IIT to GNW, assume that experiences are not observable or directly measurable when applying the theory to physical systems. That is, experiences themselves are never identified or used in obs but are rather inferred based on some dataset o that contains report or other behavioral indicators.

Next, we explore how the datasets in $\mathcal{O}$ are used to make predictions about the experience of a physical system.

Predictions

A theory of consciousness makes predictions about the experience of some physical system in some configuration, state, or dynamics, p, based on some dataset o. To this end, a theory carries within its definition a set or space E whose elements correspond to various different *conscious experiences* a system could have. The interpretation of this set varies from theory to theory, ranging from descriptions of the level of conscious experience in early versions of IIT, descriptions of the level and content of conscious experience in contemporary IIT (Kleiner and Tull 2020), or the description only of whether a presented stimuli is experienced in GNW or HOT. We sometimes refer to elements e of E simply as *experiences*.

Formally, this means that a prediction considers an experimental dataset $o \in \mathcal{O}$ (determined by *obs*) and specifies an element of the experience space E. We denote this as *pred*, for "prediction," which is a map from $\mathcal{O}$ to E. The details of how individual datasets are being used to make predictions again do not matter for the sake of our investigation. What matters is that a procedure exists, and this is captured by *pred*. However, we have to take into account that a single dataset $o \in \mathcal{O}$ may not predict only one single experience. In general, *pred* may only allow an experimenter to constrain experience of the system in that it only specifies a subset of all experiences a theory models. We denote this subset by pred(o). Thus, pred is also a correspondence

$$\text{pred} : \mathcal{O} \twoheadrightarrow E.$$

Shown in Fig. 1 is the full set of terms needed to formally define how most contemporary theories of consciousness make predictions about the experience. So far, what we have said is very general. Indeed, the force and generalizability of our

$$P \xrightarrow{\quad obs \quad} \mathcal{O} \xrightarrow{\quad pred \quad} E$$

Figure 1. We assume that an experimental setup apt for a particular model of consciousness has been chosen for some class of physical systems P, wherein $p \in P$ represents the dynamics or configurations of a particular physical system. $\mathcal{O}$ then denotes all datasets that can arise from observations or measurements on P. Measuring the observables of p maps to datasets $o \in \mathcal{O}$, which is denoted by the *obs* correspondence. E represents the mathematical description of experience given by the theory or model of consciousness under consideration. In the simplest case, this is just a set whose elements indicate whether a stimulus has been perceived consciously or not, but far more complicated structures can arise (e.g. in IIT). The correspondence *pred* describes the process of prediction as a map from $\mathcal{O}$ to E.

argument comes from the fact that we do not have to define *pred* explicitly for the various models we consider. It suffices that it exists, in some form or the other, for the models under consideration.

It is crucial to note that predicting states of consciousness alone does not suffice to test a model of consciousness. Some have previously criticized theories of consciousness, IIT in particular, just based off of their counter-intuitive predictions. An example is the criticism that relatively simply grid-like networks have high Φ (Aaronson 2014; Tononi 2014). However, debates about counter-intuitive predictions are not meaningful by themselves, since pred alone does not contain enough information to say whether a theory is true or false. The most a theory could be criticized for is either not fitting our own phenomenology or not being parsimonious enough, neither of which are necessarily violated by counter-intuitive predictions. For example, it may actually be parsimonious to assume that many physical systems have consciousness (Goff 2017). That is, speculation about acceptable predictions by theories of consciousness must implicitly rely on a comparative reference to be meaningful, and speculations that are not explicit about their reference are uninformative.

Inferences

As discussed in the previous section, a theory is unfalsifiable given just predictions alone, and so *pred* must be compared to something else. Ideally, this would be the actual conscious experience of the system under investigation. However, as noted previously, the class of theories we focus on here assumes that experience itself is not part of the observables. For this reason, the experience of a system must be inferred separately from a theory's prediction to create a basis of comparison. Most commonly, such inferences are based on *reports*. For instance, an inference might be based on an experimental participant reporting on the switching of some perceptually bistable image (Blake *et al.* 2014) or on reports about seen vs. unseen images in masking paradigms (Alais *et al.* 2010).

It has been pointed out that report in a trial may interfere with the actual isolation of consciousness, and there has recently been the introduction of so-called "no-report paradigms" (Tsuchiya *et al.* 2015). In these cases, report is first correlated to some autonomous phenomenon like optokinetic nystagmus (stereotyped eye movement), and then the experimenter can use this instead of the subject's direct reports to infer their experiences. Indeed, there can even be simpler cases where report is merely assumed: e.g., that in showing a red square, a participant will experience a red square without necessarily asking the participant since previously that participant has proved compos mentis. Similarly, in cases of nonhumans incapable of verbal report, "report" can be broadly construed as behavior or output.

All these cases can be broadly described as being a case of inference off of some data. These data might be actual reports (like a participant's button pushes) or may be based off of physiological reactions (like no-report paradigms) or may be the outputs of a neural network or set of logic gates, such as the results of an image classification task (LeCun *et al.* 2015). Therefore, the inference can be represented as a function, inf(o), between a dataset o generated by observation or measurement of the physical system, and the set of postulated experiences in the model of consciousness, E:

$$\text{inf} : \mathcal{O} \to E .$$

Defining inf as a function means that we assume that for every experimental dataset o, one single experience in E is inferred during the experiment. Here, we use a function instead of a correspondence for technical and formal ease, which does not affect our results: if two correspondences to the same space are given, one of them can be turned into a function. (If inf is a correspondence, one defines a new space E' by $E' := \{\text{inf}(o) | o \in \mathcal{O}\}$. Every individual element of this space describes exactly what can be inferred from one dataset $o \in \mathcal{O}$, so that $\text{inf}' : \mathcal{O} \to E'$ is a function. The correspondence obs is then redefined, for every $e' \in E'$, by the requirement that $e' \in \text{obs}'(o)$ iff $e \in \text{obs}(o)$ for some $e \in e'$.) The *inf* function is flexible enough to encompass both direct report, no-report, input/output analysis, and also assumed-report cases. It is a mapping that describes the process of inferring the conscious experience of a system from data recorded in the experiments. Both *inf* and *pred* are depicted in Fig. 2.

It is worth noting that we have used here the same class $\mathcal{O}$ as in the definition of the prediction mapping pred above. This makes sense because the inference process also uses data obtained in experimental trials, such as reports by a subject. So both *pred* and *inf* can be described to operate on the same total dataset measured, even though they usually use different parts of this dataset (cf. below).

Falsification

We have now introduced all elements which are necessary to formally say what a falsification of a theory of consciousness is. To falsify, a theory of consciousness requires mismatch between an experimenter's inference (generally based on report) and the predicted consciousness of the subject. In order to describe this, we consider some particular experimental trial, as well as inf and pred.

Definition 2.1. *There is a falsification at $o \in \mathcal{O}$ if we have*

$$\text{inf}(o) \notin \text{pred}(o) . \tag{2}$$

This definition can be spelled out in terms of individual components of E. To this end, for any given dataset $o \in \mathcal{O}$, let $e_r := \text{inf}(o)$ denote the experience that is being inferred, and let $e_p \in \text{obs}(o)$ be one of the experiences that is predicted based off of some dataset. Then (2) simply states that we have $e_p \neq e_r$ for all possible predictions $e_p \in \text{obs}(o)$. None of the predicted states of experience is equal to the inferred experience.

What does Equation (2) mean? There are two cases which are possible. Either, the prediction based on the theory of consciousness is correct, and the inferred experience is wrong. Or the prediction is wrong, so that in this case the model would be falsified. In short: either the prediction process or the inference process is wrong.
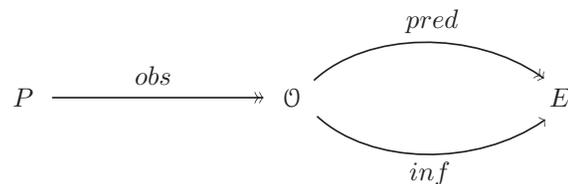

Figure 2. Two maps are necessary for a full experimental setup, one that describes a theory's predictions about experience (*pred*), another that describes the experimenter's inference about it (*inf*). Both map from a dataset $o \in \mathcal{O}$ collected in an experimental trail to some subset of experiences described by the model, E.

We remark that if there is a dataset o on which the inference procedure inf or the prediction procedure pred cannot be used, then this dataset cannot be used in falsifying a model of consciousness. Thus, when it comes to falsifications, we can restrict to datasets o for which both procedures are defined.

In order to understand in more detail what is going on if (2) holds, we have to look into a single dataset $o \in \mathcal{O}$. This will be of use later.

Generally, inf and obs will make use of different part of the data obtained in an experimental trial. For example, in the context of IIT or GNW, data about the internal structure and state of the brain will be used for the prediction. These data can be obtained from an fMRI scan or EEG measurement. The state of consciousness on the other hand can be inferred from verbal reports. Pictorially, we may represent this as in Fig. 3. We use the following notation:

o_i For a chosen dataset $o \in \mathcal{O}$, we denote the part of the dataset which is used for the prediction process by o_i (for "internal" data). This can be thought of as data about the internal workings of the system. We call o_i the *prediction data* in o.

o_r For a chosen dataset $o \in \mathcal{O}$, we denote the part of the dataset which is used for inferring the state of experience by o_r (for "report" data). We call it the *inference data* in o.

Note that in both cases, the subscript can be read similarly as the notation for restricting a set. We remark that a different kind of prediction could be considered as well, where one makes use of the inverse of pred. In Appendix B, we prove that this is in fact equivalent to the case considered here, so that Definition 2.1 indeed covers the most general situation.

Summary

In summary, for testing of a theory of consciousness we have introduced the following notion:

P denotes a class of physical systems that could have been tested, in principle, in the experiment under consideration, each in various different configurations. In most cases, every $p \in P$ thus describes a physical system in a particular state, dynamical trajectory, or configuration.

obs is a correspondence which contains all details on how the measurements are set up and what is measured. It describes how measurement results (datasets) are determined by a system configuration under investigation. This correspondence is given, though usually not explicitly known, once a choice of measurement scheme has been made.

$\mathcal{O}$ is the class of all possible datasets that can result from observations or measurements of the systems in the class P. Any single experimental trial results in a single dataset $o \in \mathcal{O}$, whose data are used for making predictions based on the theory of consciousness and for inference purposes.

pred describes the process of making predictions by applying some theory of consciousness to a dataset o. It is therefore a mapping from $\mathcal{O}$ to E.

E denotes the space of possible experiences specified by the theory under consideration. The result of the prediction is a subset of this space, denoted as pred(o). Elements of this subset are denoted by e_i and describe predicted experiences.

inf describes the process of inferring a state of experience from some observed data, e.g., verbal reports, button presses or using no-report paradigms. Inferred experiences are denoted by e_r.

The Substitution Argument

Substitutions are changes of physical systems (i.e. the substitution of one for another) that leave the inference data invariant, but may change the result of the prediction process. A specific case of substitution, the unfolding of a reentrant neural network to a feedforward one, was recently applied to IIT to argue that IIT cannot explain consciousness (Doerig *et al.* 2019).

Here, we show that, in general, the contemporary notion of falsification in the science of consciousness exhibits this fundamental flaw for almost all contemporary theories, rather than being a problem for a particular theory. This flaw is based on the independence between the data used for inferences about consciousness (like reports) and the data used to make predictions about consciousness. We discuss various responses to this flaw in Objections section.

We begin by defining what a substitution is in Substitutions section, show that it implies falsifications in Substitutions imply falsifications section and analyze the particularly problematic case of universal substitutions in Universal substitutions imply complete falsification section. In When does a universal substitution exist? section, we prove that universal
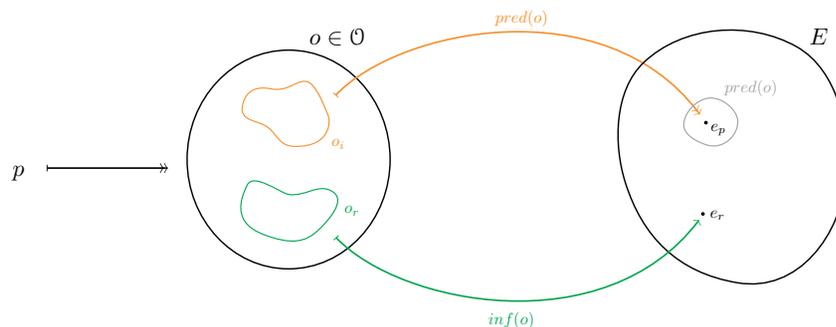

Fig. 3. This figure represents the same setup as Fig. 2. The left circle depicts one single dataset o. o_i (orange) is the part of the dataset used for prediction. o_r (green) is the part of the dataset used for inferring the state of experience. Usually the green area comprises verbal reports or button presses, whereas the orange area comprises the data obtained from brain scans. The right circle depicts the experience space E of a theory under consideration. e_p denotes a predicted experience while e_r denotes the inferred experience. Therefore, in total, to represent some specific experimental trial we use $p \in P$, $o \in \mathcal{O}$, $e_r \in E$ and $e_p \in E$, where $e_p \in$ pred(o).

substitutions exist if prediction and inference data are independent and give some examples of already-known cases.

Substitutions

In order to define formally what a substitution is, we work with the inference content o_r of a dataset o as introduced in Falsification section. We first denote the class of all physical configurations which could have produced the inference content o_r upon measurement by P_{o_r}. Using the correspondence obs which describes the relation between physical systems and measurement results, this can be defined as

$$P_{o_r} := \{p \in P | o_r \in \mathrm{obs}(p)\},\tag{3}$$

where $\mathrm{obs}(p)$ denotes all possible datasets that can be measured if the system p is under investigation and where $o_r \in \mathrm{obs}(p)$ is a shorthand for $o \in \mathrm{obs}(p)$ with inference content o_r.

Any map of the form $S : P_{o_r} \to P_{o_r}$ takes a system configuration p which can produce inference content o_r to another system's configuration $S(p)$ which can produce the same inference content. This allows us to define what a substitution is formally. In what follows, the $\bigcirc$ indicates the composition of the correspondences obs and pred to give a correspondence from P to E, which could also be denoted as $\mathrm{pred}(\mathrm{obs}(p))$ (That is, $\mathrm{pred} \bigcirc \mathrm{obs}(p) = \{e \in E | e \in \mathrm{pred}(o)$ for some $o \in \mathrm{obs}(p)\}$, it is the image under pred of the set $\mathrm{obs}(o)$.), and $\cap$ denotes the intersection of sets.

Definition 3.1. *There is a o_r-substitution if there is a transformation $S : P_{o_r} \to P_{o_r}$ such that at least for one $p \in P_{o_r}$*

$$\mathrm{pred} \bigcirc \mathrm{obs}(p) \cap \mathrm{pred} \bigcirc \mathrm{obs}(S(p)) = \varnothing.\tag{4}$$

In words, a substitution requires there to be a transformation S which keeps the inference data constant but changes the prediction of the system. So much in fact that the prediction of the original configuration p and of the transformed configuration $S(p)$ are fully incompatible, i.e. there is no single experience e which is contained in both predictions. Given some inference data o_r, an o_r-substitution then requires this to be the case for at least one system configuration p that gives this inference data. In other words, the transformation S is such that for at least one p, the *predictions change completely*, while the inference content o_r is preserved.

A pictorial definition of substitutions is given in Fig. 4. We remark that if pred and obs were functions, so that $\mathrm{pred} \bigcirc \mathrm{obs}(p)$ only contained one element, Equation (4) would be equivalent to $\mathrm{pred}(\mathrm{obs}(p)) \neq \mathrm{pred}(\mathrm{obs}(S(p)))$.

We will find below that the really problematic case arises if there is an o_r-substitution for every possible inference content o_r. We refer to this case as a universal substitution.

Definition 3.2. *There is a universal substitution if there is an o_r-substitution $S_{o_r} : P_{o_r} \to P_{o_r}$ for every o_r.*

We recall that according to the notation introduced in Falsification section, the inference content of any dataset $o \in \mathcal{O}$ is denoted by o_r (adding the subscript r). Thus, the requirement is that there is an o_r-substitution $S_{o_r} : P_{o_r} \to P_{o_r}$ for every inference data that can pertain in the experiment under consideration (for every inference data that is listed in $\mathcal{O}$). The subscript o_r of S_{o_r} indicates that the transformation S in Definition 3.1 can be chosen differently for different o_r. Definition 3.2 does not require there to be one single transformation that works for all o_r.

Substitutions imply falsifications

The force of our argument comes from the fact that if there are substitutions, then this necessarily leads to mismatches between inferences and predictions. This is shown by the following lemma.

Lemma 3.3. *If there is a o_r-substitution, there is a falsification at some $o \in \mathcal{O}$.*

Proof. Let p be the physical system in Definition 3.1 and define $p' = S(p)$. Let $o \in \mathrm{obs}(p)$ be a dataset of p which has inference content o_r and let o' be a dataset of p' which has the same inference content o_r, guaranteed to exist by the definition of P_{o_r} in (3). Equation (4) implies that

$$\mathrm{pred}(o) \cap \mathrm{pred}(o') = \varnothing.\tag{5}$$

Since, however, $o_r = o'_r$, we have $\mathrm{inf}(o) = \mathrm{inf}(o')$. Thus we have either $\mathrm{inf}(o) \notin \mathrm{pred}(o)$ or $\mathrm{inf}(o') \notin \mathrm{pred}(o')$, or both. Thus there is either a falsification at o, a falsification at o', or both. □

The last lemma shows that if there are substitutions, then there are necessarily falsifications. This might, however, not be considered too problematic, since it could always be the case that the model is right whereas the inferred experience is wrong. Inaccessible predictions are not unusual in science. A fully problematic case only pertains for universal substitutions, i.e., if there is an o_r-substitution for every inference content o_r that can arise in an experiment under consideration.

Universal substitutions imply complete falsification

In Falsification section, we have defined falsifications for individual datasets $o \in \mathcal{O}$. Using the "insight view" of single datasets, we can refine this definition somewhat by relating it to the inference content only.

Definition 3.4. *There is an o_r-falsification if there is a falsification for some $o \in \mathcal{O}$ which has inference content o_r.*

This definition is weaker than the original definition, because among all datasets which have inference content o_r, only one needs to exhibit a falsification. Using this notion, the next lemma specifies the exact relation between substitutions and falsifications.

Lemma 3.5. *If there is an o_r-substitution, there is an o_r-falsification.*

Proof. This lemma follows directly from the proof of Lemma 3.3 because the datasets o and o' used in that proof both have inference content o_r. □

This finally allows us to show our first main result. It shows that if a universal substitution exists, the theory of consciousness under consideration is falsified. We explain the meaning of this proposition after the proof.

Proposition 3.6. *If there is a universal substitution, there is an o_r-falsification for all possible inference contents o_r.*

Proof. By definition of universal substitution, there is an o_r-substitution for every o_r. Thus, the claim follows directly from Lemma 3.5. □

In combination with Definition 3.4, this proposition states that for every possible report (or any other type of inference procedure, cf. our use of terminology in Falsification section), there is a dataset o which contains the report's data and for which we have

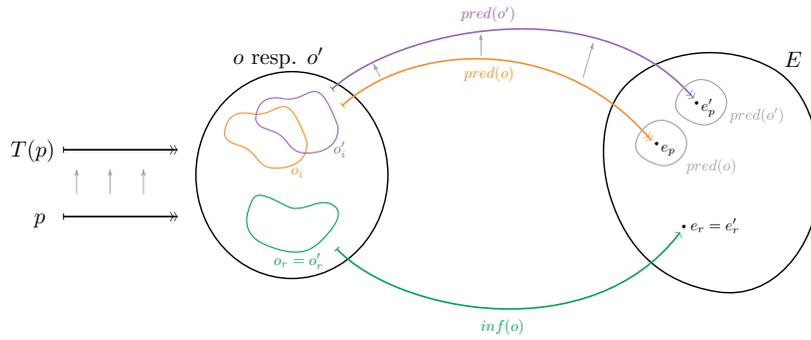$$\mathrm{inf}(o_r) \notin \mathrm{pred}(o),\tag{6}$$

Figure 4. This picture illustrates substitutions. Assume that some dataset o with inference content o_r is given. A substitution is a transformation S of physical systems which leaves the inference content o_r invariant but which changes the result of the prediction process. Thus whereas p and $S(p)$ have the same inference content o_r, the prediction content of experimental datasets is different; different in fact to such an extent that the predictions of consciousness based on these datasets are incompatible (illustrated by the nonoverlapping gray circles on the right). Here, we have used that by definition of P_{o_r}, every $\tilde{p} \in P_{o_r}$ yields at least one dataset o' with the same inference content as o and have identified o and o' in the drawing.

where we have slightly abused notation in writing $\inf(o_r)$ instead of $\inf(o)$ for clarity. This implies that one of two cases needs to pertain: either at least one of the inferred experiences $\inf(o_r)$ is correct, in which case the corresponding prediction is wrong and the theory needs to be considered falsified. The only other option is that for *all* inference contents o_r, the prediction $\text{pred}(o)$ is correct, which qua (6) implies that no single inference $\inf(o_r)$ points at the correct experience, so that the inference procedure is completely wrong. This shows that Proposition 3.6 can equivalently be stated as follows.

Proposition 3.7. *If there is a universal substitution, either every single inference operation is wrong or the theory under consideration is already falsified.*

Next, we discuss under which circumstances a universal substitution exists.

When does a universal substitution exist?

In the last section, we have seen that if a universal substitution exists, this has strong consequences. In this section, we discuss under what conditions universal substitutions exist.

Theories need to be minimally informative

We have taken great care above to make sure that our notion of prediction is compatible with incomplete or noisy datasets. This is the reason why pred is a correspondence, the most general object one could consider. For the purpose of this section, we add a gentle assumption which restricts pred slightly: we assume that every prediction carries at least a minimal amount of information. In our case, this means that for every prediction $\text{pred}(o)$, there is at least one other prediction $\text{pred}(o')$ which is different from $\text{pred}(o)$. Put in simple terms, this means that we do not consider theories of consciousness which have only a single prediction.

In order to take this into account, for every $o \in \mathcal{O}$, we define $\bar{o} := \text{obs}(\text{obs}^{-1}(o))$, which comprises exactly all those datasets which can be generated by physical systems p that also generate o. When applying our previous definitions, this can be fleshed out as

$$\bar{o} = \{o' | \exists p \text{ such that } o \in \text{obs}(p) \text{ and } o' \in \text{obs}(p)\}. \tag{7}$$

Using this, we can state our *minimal information assumption* in a way that is compatible with the general setup displayed in :

We assume that the theories of consciousness under consideration are *minimally informative* in that for every $o \in \mathcal{O}$, there exists an $o' \in \mathcal{O}$ such that

$$\text{pred}(\bar{o}) \cap \text{pred}(\bar{o}') = \varnothing. \tag{8}$$

Inference and prediction data are independent

We have already noted, that in most experiments, the prediction content o_i and inference content o_r consist of different parts of a dataset. What is more, they are usually assumed to be independent, in the sense that changes in o_i are possible while keeping o_r constant. This is captured by the next definition.

Definition 3.8. *Inference and prediction data are independent if for any o_i, o_i' and o_r, there is a variation*

$$\nu : P \to P \tag{9}$$

such that $o_i \in \text{obs}(p)$, $o_i' \in \text{obs}(\nu(p))$ *but* $o_r \in \text{obs}(p)$ *and* $o_r \in \text{obs}(\nu(p))$ *for some $p \in P$.*

Here, we use the same shorthand as in (3). For example, the requirement $o_i \in \text{obs}(p)$ is a shorthand for there being an $o \in \text{obs}(p)$ which has prediction content o_i. The variation ν in this definition is a variation in P, which describes physical systems which could, in principle, have been realized in an experiment (cf. Summary section). We note that a weaker version of this definition can be given which still implies our results below, cf. Appendix A. Note that if inference and prediction data are not independent, e.g., because they have a common cause, problems of tautologies loom large, cf. Objections section. Throughout the text, we often refer to Definition 3.8 simply as "independence."

Universal substitutions exist

Combining the last two sections, we can now prove that universal substitutions exist.

Proposition 3.9. *If inference and prediction data are independent, universal substitutions exist.*

Proof. To show that a universal substitution exists, we need to show that for every $o \in \mathcal{O}$, an o_r-substitution exists (Definition 3.1). Thus assume that an arbitrary $o \in \mathcal{O}$ is given. The minimal information assumption guarantees that there is an o' such that Equation (8) holds. As before, we denote the prediction content of o and o' by o_i and o'_i, respectively, and the inference content of o by o_r.

Since inference and prediction data are independent, there exists a $p \in P$ as well as a $\nu : P \to P$ such that $o_i \in \mathrm{obs}(p)$, $o'_i \in \mathrm{obs}(\nu(p))$, $o_r \in \mathrm{obs}(p)$ and $o_r \in \mathrm{obs}(\nu(p))$. By Definition (7), the first two of these four conditions imply that $\mathrm{obs}(p) \subset \overline{o}$ and $\mathrm{obs}(\nu(p)) \subset \overline{o}'$. Thus, Equation (8) applies and allows us to conclude that

$$\mathrm{pred}(\mathrm{obs}(p)) \cap \mathrm{pred}(\mathrm{obs}(\nu(p))) = \varnothing \,.$$

Via Equation (3), the latter two of the four conditions imply that $p \in P_{o_r}$ and $\nu(p) \in P_{o_r}$. Thus, we may restrict ν to P_{o_r} to obtain a map

$$S : P_{o_r} \to P_{o_r} \,,$$

which in light of the first part of this proof exhibits at least one $p \in P_{o_r}$ which satisfies (4). Thus we have shown that an o_r-substitution exists. Since o was arbitrary, it follows that a universal substitution exists. □

The intuition behind this proof is very simple. In virtue of our assumption that theories of consciousness need to be minimally informative, for any dataset o, there is another dataset o' which makes a nonoverlapping prediction. But in virtue of inference and prediction data being independent, we can find a variation that changes the prediction content as prescribed by o and o' but keeps the inference content constant. This suffices to show that there exists a transformation S as required by the definition of a substitution.

Combining this result with Proposition 3.7, we finally can state our main theorem.

Theorem 3.10. *If inference and prediction data are independent, either every single inference operation is wrong or the theory under consideration is already falsified.*

Proof. The theorem follows by combining Propositions 3.9 and 3.7. □

In the next section, we give several examples of universal substitutions, before discussing various possible responses to our result in Objections section.

Examples of data independence

Our main theorem shows that testing a theory of consciousness will necessarily lead to its falsification if inference and prediction data are independent (Definition 3.8), and if at least one single inference can be trusted (Theorem 3.10). In this section, we give several examples that illustrate the independence of inference and prediction data. We take report to mean output, behavior, or verbal report itself and assume that prediction data derives from internal measurements.

Artificial neural networks. ANNs, particularly those trained using deep learning, have grown increasingly powerful and capable of human-like performance (LeCun *et al.* 2015; Bojarski *et al.* 2016). For any ANN, report (output) is a function of node states. Crucially, this function is noninjective, i.e., some nodes are not part of the output. For example, in deep learning, the report is typically taken to consist of the last layer of the ANN, while the hidden layers are not taken to be part of the output. Correspondingly, for any given inference data, one can construct a ANN with arbitrary prediction data by adding nodes, changing connections and changing those nodes which are not part of the output. Put differently, one can always substitute a given ANN with another with different internal observables but identical or near-identical reports. From a mathematical perspective, it is well-known that both feedforward ANNs and recurrent ANNs can approximate any given function (Hornik *et al.* 1989; Schäfer and Zimmermann 2007). Since reports are just some function, it follows that there are viable universal substitutions.

A special case thereof is the unfolding transformation considered in Doerig *et al.* (2019) in the context of IIT. The arguments in this article constitute a proof of the fact that for ANNs, inference and prediction data are independent (Definition 3.8). Crucially, our main theorem shows that this has implications for all minimally informative theories of consciousness. A similar result (using a different characterization of theories of consciousness than minimally informative) has been shown in Kleiner (2020).

Universal computers. Turing machines are extremely different in architecture than ANNs. Since they are capable of universal computation (Turing 1937), they should provide an ideal candidate for a universal substitution. Indeed, this is exactly the reasoning behind the Turing test of conversational artificial intelligence (Turing 1950). Therefore, if one believes it is possible for a sufficiently fast Turing machine to pass the Turing test, one needs to accept that substitutions exist. Notably, Turing machines are just one example of universal computation, and there are other instances of different parameter spaces or physical systems that are capable thereof, such as cellular automata (Wolfram 1984).

Universal intelligences. There are models of universal intelligence that allow for maximally intelligent behavior across any set of tasks (Hutter 2003). For instance, consider the AIXI model, the gold-standard for universal intelligence, which operates via Solomonoff induction (Solomonoff 1964; Hutter 2004). The AIXI model generates an optimal decision making over some class of problems, and methods linked to it have already been applied to a range of behaviors, such as creating "AI physicists" (Wu and Tegmark 2019). Its universality indicates it is a prime candidate for universal substitutions. Notably, unlike a Turing machine, it avoids issues of precisely how it is accomplishing universal substitution of report, since the algorithm that governs the AIXI model behavior is well-described and "relatively" simple.

The above are all real and viable classes of systems that are used everyday in science and engineering which all provide different viable universal substitutions if inferences are based on reports or outputs. They show that in normal experimental setups such as the ones commonly used in neuroscientific research into consciousness (Frith *et al.* 1999), inference and prediction data are indeed independent, and dependency is not investigated nor properly considered. It is always possible to substitute the physical system under consideration with another that has different internal observables, and therefore different predictions, but similar or identical reports. Indeed, recent research in using the work introduced in this work shows that even different spatiotemporal models of a system can be substituted for one another, leading to falsification (Hanson and Walker 2020). We have not considered possible but less reason-

able examples of universal substitutions, like astronomically large look-up ledgers of reports.

As an example of our Main Theorem 3.10, we consider the case of IIT. Since the theory is normally applied in Boolean networks, logic gates, or artificial neural networks, one usually takes report to mean "output." In this case, it has already been proven that systems with different internal structures and hence different predicted experiences, can have identical input/output (and therefore identical reports or inferences about report) (Albantakis and Tononi 2019). To take another case: within IIT it has already been acknowledged that a Turing machine may have a wildly different predicted contents of consciousness for the same behavior or reports (Koch 2019). Therefore, data independence during testing has already been shown to apply to IIT under its normal assumptions.

Inference and Prediction Data Are Strictly Dependent

An immediate response to our main result showing that many theories suffer from *a priori* falsification would be to claim that it offers support of theories which define conscious experience in terms of what is accessible to report. This is the case, e.g., for behaviorist theories of consciousness but might arguably also be the case for some interpretations of global workspace theory or fame in the brain proposals. In this section, we show that this response is not valid, as theories of this kind, where inference and prediction data are strictly dependent, are unfalsifiable.

In order to analyze this case, we first need to specifically outline how theories can be pathologically unfalsifiable. Clearly, the goal of the scientific study as a whole is to find, eventually, a theory of consciousness that are empirically adequate and therefore corroborated by *all* experimental evidence. Therefore, not being falsified in experiments is a necessary condition (though not sufficient) any purportedly "true" theory of consciousness needs to satisfy. Therefore, not being falsifiable by the set of possible experiments per se is not a bad thing. We seek to distinguish this from cases of unfasifiability due to pathological assumptions that underlie a theory of consciousness, assumptions which render an experimental investigation meaningless. Specifically, a pathological dependence between inferences and predictions leads to theories which are unfalsifiable.

Such unfalsifiable theories can be identified neatly in our formalism. To see how, recall that $\mathcal{O}$ denotes the class of all datasets that can result from an experiment investigating the physical systems in the class P. Put differently, it contains all datasets that could, in principle, appear when probed in the experiment. This is *not* the class of all possible datasets of type $\mathcal{O}$ one can think of. Many datasets which are of the same form as elements of $\mathcal{O}$ might simply not arise in the experiment under consideration. We denote the class of all possible datasets as:

$$\overline{\mathcal{O}} : \text{All possible data sets of type } \mathcal{O}.$$

Intuitively, in terms of possible worlds semantics, $\mathcal{O}$ describes the datasets which could appear, for the type of experiment under consideration, in the actual world. $\overline{\mathcal{O}}$, in contrast, describes the datasets which could appear in this type of experiment in any possible world. For example, $\overline{\mathcal{O}}$ contains datasets which can only occur if consciousness attaches to the

physical in a different way than it actually does in the actual word.

By construction, $\mathcal{O}$ is a subset of $\overline{\mathcal{O}}$, which describes which among the possible datasets actually arises across experimental trials. Hence, $\mathcal{O}$ also determines which theory of consciousness is compatible with (i.e. not falsified by) experimental investigation. However, $\overline{\mathcal{O}}$ defines all possible datasets independent of any constraint by real empirical results, i.e., all possible imaginable datasets.

Introduction of $\overline{\mathcal{O}}$ allows us to distinguish the pathological cases of unfalsifiability mentioned above. Whereas any purportedly true theory should only fail to be falsified with respect to the experimental data $\mathcal{O}$, a pathological unfalsifiability pertains if a theory cannot be falsified at all, i.e. over $\overline{\mathcal{O}}$. This is captured by the following definition.

Definition 4.1. *A theory of consciousness which does not have a falsification over $\overline{\mathcal{O}}$ is empirically unfalsifiable.*

Here, we use the term "empirically unfalsifiable" to highlight and refer to the pathological notion of unfalsifiability. Intuitively speaking, a theory which satisfies this definition appears to be true independently of any experimental investigation, and without the need for any such investigation. Using $\overline{\mathcal{O}}$, we can also define the notion of strict dependence in a useful way.

Definition 4.2. *Inference and prediction data are strictly dependent if there is a function f such that for any $o \in \overline{\mathcal{O}}$, we have $o_i = f(o_r)$.*

This definition says that there exists a function f which for every possible inference data o_r allows to deduce the prediction data o_i. We remark that the definition refers to $\overline{\mathcal{O}}$ and not $\mathcal{O}$, as the dependence of inference and prediction considered here holds by assumption and is not simply asserting a contingency in nature.

The definition is satisfied, e.g., if inference data is equal to prediction data, i.e., if $o_i = o_r$, where f is simply the identity. This is the case, e.g., for behaviorist theories (Skinner 1938) of consciousness, where consciousness is equated directly with report or behavior, or for precursors of functionalist theories of consciousness that are based on behavior or input/output (Putnam 1960). The definition is also satisfied in the case where prediction data are always a subset of the inference data:

$$o_i \subseteq o_r . \tag{10}$$

Here, f is simply the restriction function. This arguably applies to global workspace theory (Baars 2005), the "attention schema" theory of consciousness (Graziano and Webb 2015) or "fame in the brain" (Dennett 1991) proposals.

In all these cases, consciousness is generated by—and hence needs to be predicted via—what is accessible to report or output. In terms of Block's distinction between phenomenal consciousness and access consciousness (Block 1996), Equation (10) holds true whenever a theory of consciousness is under investigation where access consciousness determines phenomenal consciousness.

Our second main theorem is the following.

Theorem 4.3. *If a theory of consciousness implies that inference and prediction data are strictly dependent, then it is either already falsified or empirically unfalsifiable.*

Proof. To prove the theorem, it is useful to consider the inference and prediction content of datasets explicitly. The possible pairings that can occur in an experiment are given by

$$\mathcal{O}_{\exp} := \{(o_i, o_r) \mid o \in \mathcal{O}\}, \tag{11}$$

where we have again used our notation that o_i denotes the prediction data of o, and similar for o_r. To define the possible pairings that can occur in $\overline{\mathcal{O}}$, we let $\mathcal{O}_i$ denote the class of all prediction contents that arise in $\mathcal{O}$, and $\mathcal{O}_r$ denote the class of all inference contents that arise in $\mathcal{O}$. The set of all conceivable pairings is then given by

$$\mathcal{O}_{\text{all}} := \{(o_i, o_r') \mid o \in \mathcal{O}, o' \in \mathcal{O}\} \tag{12}$$

$$= \{(o_i, o_r') \mid o_i \in \mathcal{O}_i, o_r' \in \mathcal{O}_r\}. \tag{13}$$

Crucially, here, o_i and o_r' do not have to be part of the same dataset o. Combined with Definition 2.1, we conclude that there is a falsification over $\overline{\mathcal{O}}$ if for some $(o_i, o_r') \in \mathcal{O}_{\text{all}}$, we have $\inf(o) \notin \text{pred}(o')$, and there is a falsification over $\mathcal{O}$ if for some $(o_i, o_r) \in \mathcal{O}_{\exp}$, we have $\inf(o) \notin \text{pred}(o)$.

Next we show that if inference and prediction data are strictly dependent, then $\mathcal{O}_{\text{all}} = \mathcal{O}_{\exp}$ holds. We start with the set $\mathcal{O}_{\text{all}}$ as defined in (12). Expanding this definition in words, it reads

$$\mathcal{O}_{\text{all}} = \{(d_i, d_r) \mid \exists o \in \mathcal{O} \text{ such that } d_r = o_r \text{ and } \exists \bar{o} \in \mathcal{O} \text{ such that } d_i = \bar{o}_i\}, \tag{14}$$

where we have symbols d_i and d_r to denote prediction and inference data independently of any dataset o.

Assume that the first condition in this expression, $d_r = o_r$ holds for some $o \in \mathcal{O}$. Since inference and prediction data are strictly dependent, we have $d_i = f(d_r)$. Furthermore, for the same reason, the prediction content o_i of the dataset o satisfies $o_i = f(o_r)$. Applying the function f to both sides of the first condition gives $f(d_r) = f(o_r)$, which thus in turn implies $o_i = d_i$. This means that the o that satisfies the first condition in (14) automatically also satisfies the second condition. Therefore, due to inference and prediction data being strictly dependent, (14) is equivalent to

$$\mathcal{O}_{\text{all}} = \{(d_i, d_r) \mid \exists o \in \mathcal{O} \text{ such that } d_r = o_r \text{ and } d_i = o_i\}. \tag{15}$$

This, however, is exactly $\mathcal{O}_{\exp}$ as defined in (11). Thus we conclude that if inference and prediction data are strictly dependent, $\mathcal{O}_{\text{all}} = \mathcal{O}_{\exp}$ necessarily holds.

Returning to the characterization of falsification in terms of $\mathcal{O}_{\exp}$ and $\mathcal{O}_{\text{all}}$ above, what we have just found implies that there is a falsification over $\mathcal{O}$ if and only if there is a falsification over $\overline{\mathcal{O}}$. Thus either there is a falsification over $\mathcal{O}$, in which case the theory is already falsified or there is no falsification over $\overline{\mathcal{O}}$, in which case the theory under consideration is empirically unfalsifiable. □

The gist of this proof is that if inference and prediction data are strictly dependent, then as far as the inference and prediction contents go, $\mathcal{O}$ and $\overline{\mathcal{O}}$ are the same. That is, the experiment does not add *anything* to the evaluation of the theory. It is sufficient to know only all possible datasets to decide whether there is a falsification. In practise, this would mean that knowledge of the experimental design (which reports are to be collected, on the one hand, which possible data a measurement device can produce, one the other) is sufficient to evaluate the theory, which is clearly at odds with the role of empirical evidence required in any scientific investigation. Thus, such theories are empirically unfalsifiable.

To give an intuitive example of the theorem, let us examine a theory that uses the information accessible to report in a system to predict conscious experience (perhaps this information is "famous" in the brain or is within some accessible global workspace). In terms of our notation, we can assume that o_r denotes everything that is accessible to report, and o_i denotes that part which is used by the theory to predict conscious experience. Thus, in this case we have $o_i \subseteq o_r$. Since the predicted contents are always part of what can be reported, there can never be any mismatch between reports and predictions. However, this is not only the case for $\mathcal{O}_{\exp}$ but also, in virtue of the theory's definition, for all possible datasets, i.e., $\mathcal{O}_{\text{all}}$. Therefore, such theories are empirically unfalsifiable. Experiments add no information to whether the theory is true or not, and such theories are empirically uninformative or tautological.

Objections

In this section, we discuss a number of possible objections to our results.

Restricting inferences to humans only

The examples given in section 3.4.4 show that data independence holds during the usual testing setups. This is because prima facie it seems reasonable to base inferences either on report capability or intelligent behavior in a manner agnostic of the actual physical makeup of the system. Yet this entails independence, so in these cases our conclusions apply.

One response to our results might be to restrict all testing of theories of consciousness solely to humans. In our formalisms, this is equivalent to making the strength of inferences based not on reports themselves but on an underlying biological homology. Such an *inf* function may still pick out specific experiences via reports, but the weight of the inference is carried by homology rather than report or behavior. This would mean that the substitution argument does not significantly affect consciousness research, as reports of nonhuman systems would simply not count. Theories of consciousness, so this idea goes, would be supported by abductive reasoning from testing in humans alone.

Overall, there are strong reasons to reject this restriction of inferences. One significant issue is that this objection is equivalent to saying that reports or behavior in nonhumans carry *no* information about consciousness, an incredibly strong claim. Indeed, this is highly problematic for consciousness research which often uses nonhuman animal models (Boly *et al.* 2013). For instance, cephalopods are among the most intelligent animals yet are quite distant on the tree of life from humans and have a distinct neuroanatomy, and still are used for consciousness research (Mather 2008). Even in artificial intelligence research, there is increasing evidence that deep neural networks produced brain-like structures such as grid cells, shape tuning, and visual illusions, and many others (Richards *et al.* 2019). Given these similarities, it becomes questionable why the strength of inferences should be based on homology instead of capability of report or intelligence.

What is more, restricting inferences to humans alone is unlikely to be sufficient to avoid our results. Depending on the theory under consideration, data independence might exist just in human brains alone. That is, it is probable that there are

transformations (as in Equation (9)) available within the brain wherein o_r is fixed but o_i varies. This is particularly true once one allows for interventions on the human brain by experimenters, such as perturbations like transcranial magnetic stimulation, which is already used in consciousness research (Rounis *et al.* 2010; Napolitani *et al.* 2014).

For these reasons this objection does not appear viable. At minimum, it is clear that if the objection were taken seriously, it would imply significant changes to consciousness research which would make the field extremely restricted with strong *a priori* assumptions.

Reductio ad absurdum

Another hypothetical objection to our results is to argue that they could just as well be applied to scientific theories in other fields. If this turned out to be true, this would not imply our argument is necessarily incorrect. But, the fact that other scientific theories do not seem especially problematic with regard to falsification would generate the question of whether some assumption is illegitimately strong. In order to address this, we explain which of our assumptions is specific to theories of consciousness and would not hold when applied to other scientific theories. Subsequently, we give an example to illustrate this point.

The assumption in question is that $\mathcal{O}$, the class of all datasets that can result from observations or measurements of a system, is determined by the physical configurations in P alone. That is, every single dataset o, including both its prediction content o_i and its inference content o_r, is determined by p, and *not* by a conscious experience in E. In Fig. 2, this is reflected in the fact that there is an arrow from P to $\mathcal{O}$, but no arrow from E to $\mathcal{O}$.

This assumption expresses the standard paradigm of testing theories of consciousness in neuroscience, according to which both the data used to predict a state of consciousness and the reports of a system are determined by its physical configuration alone. This, in turn, may be traced back to consciousness' assumed subjective and private nature, which implies that any empirical access to states of consciousness in scientific investigations is necessarily mediated by a subject's reports, and to general physicalist assumptions.

This is different from experiments in other natural sciences. If there are two quantities of interest whose relation is to be modeled by a scientific theory, then in all reasonable cases there are two *independent* means of collecting information relevant to a test of the theory, one providing a dataset that is determined by the first quantity, and one providing a dataset that is determined by the second quantity.

Consider, as an example, the case of temperature T and its relation to microphysical states. To apply our argument, the temperature T would replace the experience space E and p would denote a microphysical configuration. In order to test any particular theory about how temperature is determined by microphysical states, one would make use of two different measurements. The first measurement would access the microphysical states and would allow measurement of, say, the mean kinetic energy (if that's what the theory under consideration utilizes). This first measurement would provide a dataset o_m that replaces the prediction data o_i above. For the second measurement, one would use a thermometer or some other measuring device to obtain a dataset o_t that replaces our inference data o_r above. Comparison of the inferred temperature with the temperature that is predicted based on o_m would allow testing of the theory under consideration. These independent means

provide independent access to each of the two datasets in question. Combining o_m and o_t in one dataset o, the diagrammatic representation is

$$P \rightarrow \mathcal{O} \leftarrow T\,,$$

which differs from the case of theories of consciousness considered here, wherein the physical system determines both datasets.

Theories could be based on phenomenology

Another response to the issue of independence/dependence identified here is to propose that a theory of consciousness may not have to be falsified but can be judged by other characteristics. This is reminiscent of ideas put forward in connection with String Theory, which some have argued can be judged by elegance or parsimony alone (Carroll 2018).

In addition to elegance and parsimony, in consciousness science, one could in particular consider a theory's fit with phenomenology, i.e., how well a theory describes the general structure of conscious experience. Examples of theories that are constructed based on a fit with phenomenology are recent versions of IIT (Oizumi *et al.* 2014) or any view that proposes developing theories based on isomorphisms between the structure of experiences and the structure of physical systems or processes (Tsuchiya *et al.* 2019).

It might be suggested that phenomenological theories might be immune to aspects of the issues we outline in our results (Negro 2020). We emphasize that in order to avoid our results, and indeed the need for any experimental testing at all, a theory constructed from phenomenology has to be *uniquely derivable* from conscious experience. However, to date, no such derivation exists, as phenomenology seems to generally underdetermine the postulates of IIT (Bayne 2018; Barrett and Mediano 2019), and because it is unknown what the scope and nature of nonhuman experience is. Therefore, theories based on phenomenology can only confidently identify systems with human-like conscious experiences and cannot currently do so uniquely. Thus they cannot avoid the need for testing.

As long as no unique and correct derivation exists across the space of possible conscious experiences, the use of experimental tests to assess theories of consciousness, and hence our results, cannot be avoided.

Rejecting falsifiability

Another response to our findings might be to deny the importance of falsifications within the scientific methodology. Such responses may reference a Lakatosian conception of science, according to which science does not proceed by discarding theories immediately upon falsification, but instead consists of *research programs* built around a family of theories (Lakatos 1980). These research programs have a *protective belt* which consists of nonessential assumptions that are required to make predictions, and which can easily be modified in response to falsifications, as well as a *hard core* that is immune to falsifications. Within the Lakatosian conception of science research programs are either progressive or degenerating based on whether they can "anticipate theoretically novel facts in its growth" or not (Lakatos 1980).

It is important to note, however, that Lakatos does not actually break with falsificationism. This is why Lakatos description of science is often called "refined falsificationism" in philosophy

of science (Radnitzky 1991). Thus cases of testing theories' predictions remain relevant in a Lakatosian view in order to distinguish between progressive and degenerating research programs. Therefore, our results generally translate into this view of scientific progress. In particular, Theorem 3.10 shows that for every single inference procedure that is taken to be valid, there exists a system for which the theory makes a wrong prediction. This implies necessarily that a research program is degenerating. That is, independence implies that there is always an available substitution that can falsify any particular prediction coming from the research program.

Conclusion

In this article, we have subjected the usual scheme for testing theories of consciousness to a thorough formal analysis. We have shown that there appear to be deep problems inherent in this scheme which need to be addressed.

Crucially, in contrast to other similar results (Doerig *et al.* 2019), we do not put the blame on individual theories of consciousness, but rather show that a key assumption that is usually being made is responsible for the problems: an experimenter's inference about consciousness and a theory's predictions are generally implicitly assumed to be independent during testing across contemporary theories. As we formally prove, if this independence holds, substitutions or changes to physical systems are possible that falsify any given contemporary theory. Whenever there is an experimental test of a theory of consciousness on some physical system which does not lead to a falsification, there necessary exists another physical system which, if it had been tested, would have produced a falsification of that theory. We emphasize that this problem does not only affect one particular type of theory, e.g., those based on causal interactions like IIT; theorems apply to all contemporary neuroscientific theories of consciousness if independence holds.

In the second part of our results, we examine the case where independence does not hold. We show that if an experimenter's inferences about consciousness and a theory's predictions are instead considered to be strictly dependent, empirical unfalsifiability follows, which renders any type of experiment to test a theory uninformative. This affects all theories wherein consciousness is predicted off of reports or behavior (such as behaviorism), theories based off of input/output functions, and also theories that equate consciousness with on accessible or reportable information.

Thus, theories of consciousness seem caught between Scylla and Charybdis, requiring delicate navigation. In our opinion, there may only be two possible paths forward to avoid these dilemmas, which we briefly outline below. Each requires a revision of the current scheme of testing or developing theories of consciousness.

Lenient dependency

When combined, our main theorems show that both independence and strict dependence of inference and prediction data are problematic and thus neither can be assumed in an experimental investigation. This raises the question of whether there are reasonable cases where inference and prediction are dependent, but not strictly dependent.

A priori, in the space of possible relationships between inference and prediction data, there seems to be room for relationships that are neither independent (The substitution argument section) nor strictly dependent (Inference and prediction data are strictly dependent section). We define this relationships of

this kind as cases of *lenient dependency*. No current theory or testing paradigm that we know of satisfies this definition. Yet cases of lenient dependency cannot be excluded to exist. Such cases would technically not be beholden to either Theorem 3.10 or Theorem 4.3.

There seems to be two general possibilities of how lenient dependencies could be built. On the one hand, one could hope to find novel forms of inference that allow to surpass the problems we have identified here. This would likely constitute a major change in the methodologies of experimental testing of theories of consciousness. On the other hand, another possibility to attain lenient dependence would be to construct theories of consciousness which yield prediction functions that are designed to explicitly have a leniently dependent link to inference functions. This would likely constitute a major change in constructing theories of consciousness.

Physics is not causally closed

Another way to avoid our conclusion is to only consider theories of consciousness which do not describe the physical as causally closed (Kim 1998). That is, the presence or absence of a particular experience itself would have to make a difference to the configuration, dynamics, or states of physical systems above and beyond what would be predicted with just information about the physical system itself. If a theory of consciousness does not describe the physical as closed, a whole other range of predictions are possible: predictions which concern the physical domain itself, e.g., changes in the dynamics of the system which depend on the dynamics of conscious experience. These predictions are not considered in our setup and may serve to test a theory of consciousness without the problems we have explored here.

Supplementary Materials

(A) Weak Independence

In this section, we show how Definition 3.8 can be substantially relaxed while still ensuring our results to hold. To this end, we need to introduce another bit of formalism: We assume that predictions can be compared to establish how different they are. This is the case, e.g., in IIT where predictions map to the space of maximally irreducible conceptual structures (MICS), sometimes also called the space of Q-shapes, which carries a distance function analogous to a metric (Kleiner and Tull, 2020). We assume that for any given prediction, one can determine which of all those predictions that do not overlap with the given one is *most similar* to the latter, or equivalently which is *least different*. We calls this a *minimally differing prediction* and use it to induce a notion of *minimally differing datasets* below. Uniqueness is not required.

Let an arbitrary dataset $o \in \mathcal{O}$ be given. The minimal information assumption from Theories need to be minimally informative section ensures that there is at least one dataset o' such that Equation (8) holds. For what follows, let $o^\perp$ denote the set of all datasets which satisfy Equation (8) with respect to o,

$$o^\perp := \{ o' \in \mathcal{O} \,|\, \mathrm{pred}(\overline{o}) \cap \mathrm{pred}(\overline{o}') = \varnothing \}. \tag{16}$$

Thus $o^\perp$ contains all datasets whose prediction completely differs from the prediction of o.

Definition A.1. *We denote by* $\min(o)$ *those datasets in* $o^\perp$ *whose prediction is least different from the prediction of* o.

In many cases min(o) will only contain one dataset, but here we treat the general case where this is not so. We emphasize that the minimal information assumption guarantees that min(o) exists. We can now specify a much weaker version of Definition 3.8.

DefinitionA.2. Inference and prediction data are independent *if for any $o \in \mathcal{O}$ and $o' \in \min(o)$, there is a variation*

$$\nu : P \to P \tag{17}$$

such that $o_i \in \mathrm{obs}(p)$, $o_i' \in \mathrm{obs}(\nu(p))$ *but* $o_r \in \mathrm{obs}(p)$ *and* $o_r \in \mathrm{obs}(\nu(p))$ *for some* $p \in P$.

The difference between Definitions A.2 and 3.8 is that for a given $o \in \mathcal{O}$, the latter requires the transformation ν to exist for any $o' \in \mathcal{O}$, wheres the former only requires it to exist for minimally different datasets $o' \in \min(o)$. The corresponding proposition is the following.

Proposition A.3. *If inference and prediction data are weakly independent, universal substitutions exist.*

Proof. To show that a universal substitution exists, we need to show that for every $o \in \mathcal{O}$, an o_r-substitution exists (Definition 3.1). Thus assume that an arbitrary $o \in \mathcal{O}$ is given and pick an $o' \in \min(o)$. As before, we denote the prediction content of o and o' by o_i and o_i', respectively, and the inference content of o by o_r.

Since inference and prediction data are weakly independent, there exists a $p \in P$ as well as a $\nu : P \to P$ such that $o_i \in \mathrm{obs}(p)$, $o_i' \in \mathrm{obs}(\nu(p))$, $o_r \in \mathrm{obs}(p)$ and $o_r \in \mathrm{obs}(\nu(p))$. By Definition (7), the first two of these four conditions imply that $\mathrm{obs}(p) \subset \bar{o}$ and $\mathrm{obs}(\nu(p)) \subset \bar{o}'$. Since o' is in particular an element of $o^\perp$, Equation (8) applies and allows us to conclude that

$$\mathrm{pred}(\mathrm{obs}(p)) \cap \mathrm{pred}(\mathrm{obs}(\nu(p))) = \varnothing\,.$$

Via Equation (3), the latter two of the four conditions imply that $p \in P_{o_r}$ and $\nu(p) \in P_{o_r}$. Thus we may restrict ν to P_{o_r} to obtain a map

$$S : P_{o_r} \to P_{o_r}\,,$$

which in light of the first part of this proof exhibits at least one $p \in P_{o_r}$ which satisfies (4). Thus we have shown that an o_r-substitution exists. Since o was arbitrary, it follows that a universal substitution exists. □

The following theorem shows that Definition A.2 is sufficient to establish the claim of Theorem 3.10.

Theorem A.4. *If inference and prediction data are weakly independent, either every single inference operation is wrong or the theory under consideration is already falsified.*

Proof. The theorem follows by combining Propositions A.3 and 3.7. □

(B) Inverse Predictions

When defining falsification, we have considered predictions that take as input data about the physical configuration of a system and yield as output a state of consciousness. An alternative would be to consider the inverse procedure: a prediction which takes as input a reported stated of consciousness and yields as output some constraint on the physical configuration of the
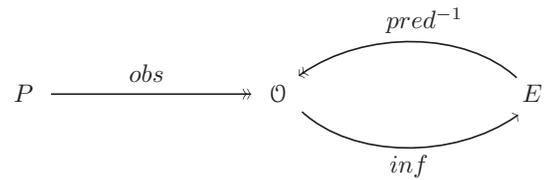

Fig. 5. The case of an inverse prediction. Rather than comparing the inferred and predicted state of consciousness, one predicts the physical configuration of a system based on the system's report and compares this with measurement results.

system that is having the conscious experience. In this section, we discuss the second case in detail.

As before, we assume that some dataset o has been measured in an experimental trail, which contains both the inference data o_r (which includes report and behavioral indicators of consciousness used in the experiment under consideration) as well as some data o_i that provides information about the physical configuration of the system under investigation. For simplicity, we will also call this *prediction data* here. Also as before, we take into account that the state of consciousness of the system has to be inferred from o_r, and again denote this inference procedure by inf.

The theory under consideration provides a correspondence $\mathrm{pred} : \mathcal{O} \twoheadrightarrow E$ which describes the process of predicting states of consciousness mentioned above. If we ask which physical configurations are compatible with a given state e of consciousness, this is simply the preimage $\mathrm{pred}^{-1}(e)$ of e under pred, defined as

$$\mathrm{pred}^{-1}(e) = \{o \in \mathcal{O}|e \in \mathrm{pred}(o)\}\,. \tag{18}$$

Accordingly, the class of all prediction data which is compatible with the inferred experience inf(o) is

$$\mathrm{pred}^{-1}(\mathrm{inf}(o))\,, \tag{19}$$

depicted in Fig. 5, and a falsification occurs in case the observed o has a prediction content o_i which is not in this set. Referring to the previous definition of falsification as *type-1* (Definition 2.1), we define this new form of falsification as *type-2*.

DefinitionB.1. *There is a type-2 falsification at $o \in \mathcal{O}$ if we have*

$$o \notin \mathrm{pred}^{-1}(\mathrm{inf}(o))\,. \tag{20}$$

In terms of the notion introduced in Summary section, Equation (20) could equivalently be written as $o_i \notin \mathrm{pred}^{-1}(\mathrm{inf}(o_r))_i$. The following lemma shows that there is a type-2 falsification if and only if there is a type-1 falsification. Hence all of our previous results apply as well to type-2 falsifications.

Lemma B.2. There is a type-2 falsification at o if and only if there is a type-1 falsification at o.

Proof. Equation (18) implies that $o \notin \mathrm{pred}^{-1}(e)$ if and only if $e \notin \mathrm{pred}(o)$. Applied to $e = \mathrm{inf}(o)$, this implies:

$$o \notin \mathrm{pred}^{-1}(\mathrm{inf}(o)) \quad \text{if and only if} \quad \mathrm{inf}(o) \notin \mathrm{pred}(o)\,.$$

The former is the definition of a type-2 falsification. The latter is Equation (2) in the definition of a type-1 falsification. Hence the claim follows. □

Acknowledgments

References

Aaronson S. Why I am not an integrated information theorist (or, the unconscious expander). In: *Shtetl-Optimized: The Blog of Scott Aaronson.* eds. Radin D., Richard D., & Karim T., 2014.

Alais D, Cass J, O'Shea RP, *et al.* Visual sensitivity underlying changes in visual consciousness. *Curr Biol* 2010;**20**: 1362–7.

Albantakis L, Tononi G. Causal composition: Structural differences among dynamically equivalent systems. *Entropy* 2019;**21**: 989.

Baars BJ. In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *J Conscious Stud* 1997;**4**:292–309.

Baars BJ. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progr Brain Res* 2005;**150**:45–53.

Ball P. Neuroscience readies for a showdown over consciousness ideas. *Quanta Mag* 2019;6.

Barrett AB, Mediano PA. The phi measure of integrated information is not well-defined for general physical systems. *J Conscious Stud* 2019;**26**:11–20.

Bayne TJ On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci Conscious* 2018;**4**: 1–8.

Blake R, Brascamp J, Heeger DJ. Can binocular rivalry reveal neural correlates of consciousness? *Philos Trans R Soc B* 2014;**369**: 20130211.

Block N. How can we find the neural correlate of consciousness? *Trends Neurosci* 1996;**19**:456–9.

Bojarski M, Del Testa D, Dworakowski D, *et al.* End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.

Boly M, Seth AK, Wilke M, *et al.* Consciousness in humans and non-human animals: recent advances and future directions. *Front Psychol* 2013;**4**:625.

Carroll SM. Beyond falsifiability: normal science in a multiverse. In: *Why Trust a Theory?* 2018, 300.

Cerullo MA. The problem with phi: a critique of integrated information theory. *PLoS Comput Biol* 2015;**11**:e1004286.

Chang AY, Biehl M, Yu Y, *et al.* Information closure theory of consciousness. Frontiers in Psychology 2020;11.

Clark A. Consciousness as generative entanglement. *J Philos* 2019;**116**:645–62.

Crick F. *Astonishing Hypothesis: The Scientific Search for the Soul.* New York: Simon and Schuster, 1994.

Crick F, Koch C. Towards a neurobiological theory of consciousness. In: *Seminars in the Neurosciences*, Vol. **2**. Saunders Scientific Publications, 1990, 263–75.

Crick FC, Koch C. What is the function of the claustrum? *Philos Trans R Soc B* 2005;**360**:1271–9.

Dehaene S, Changeux J-P. Neural mechanisms for access to consciousness. *Cogn Neurosci* 2004;**3**:1145–58.

Dehaene S, Changeux J-P. Experimental and theoretical approaches to conscious processing. *Neuron* 2011;**70**:200–27.

Del Cul A, Baillet S, Dehaene S. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol* 2007;**5**:e260.

Dennett DC. *Consciousness Explained*. Boston: Little, Brown and Co, 1991.

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.

Dołęga K, Dewhurst JE. Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese* 2020;1–26.

Frith C, Perry R, Lumer E. The neural correlates of conscious experience: an experimental framework. *Trends Cogn Sci* 1999;**3**: 105–14.

Gilmore RO, Diaz MT, Wyble BA, *et al.* Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Ann N Y Acad Sci* 2017;**1396**:5.

Goff P. *Consciousness and Fundamental Reality.* Oxford University Press, 2017.

Gosseries O, Di H, Laureys S, *et al.* Measuring consciousness in severely damaged brains. *Annu Rev Neurosci* 2014;**37**:457–78.

Graziano MS, Webb TW. The attention schema theory: a mechanistic account of subjective awareness. *Front Psychol* 2015;**6**: 500.

Hanson JR, Walker SI. Formalizing falsification of causal structure theories for consciousness across computational hierarchies. arXiv preprint arXiv:2006.07390, 2020.

HobsonAJ, *et al.* Consciousness, dreams, and inference: the Cartesian theatre revisited. *J Conscious Stud* 2014;**21**:6–32.

Hohwy J. Attention and conscious perception in the hypothesis testing brain. *Front Psychol* 2012;**3**:96.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* 1989;**2**: 359–66.

Hutter M. A gentle introduction to the universal algorithmic agent {AIXI}. In *Artificial General Intelligence*, eds. Goertzel B., and Pennachin C., Springer. 2003.

Hutter M. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin: Springer Science & Business Media, 2005.

Kim J. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge: MIT Press, 1998.

Kleiner J. Brain states matter. a reply to the unfolding argument. *Conscious Cogn* 2020;**85**:102981.

Kleiner J, Tull S. The mathematical structure of integrated information theory. arXiv preprint arXiv:2002.07655, 2020.

Koch C. *The Feeling of Life Itself: Why Consciousness is Widespread but can't be Computed*. Cambidge MA Boston: MIT Press, 2019.

Koch C, Massimini M, Boly M, *et al.* Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci* 2016;**17**: 307.

Lakatos I. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*, Vol. **1**. London: Cambridge University Press, UK, 1980.

Lamme VA. Towards a true neural stance on consciousness. *Trends Cogn Sci* 2006;**10**:494–501.

Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* 2011;**15**:365–73.

LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**: 436–44.

Massimini M, Ferrarelli F, Huber R, *et al.* Breakdown of cortical effective connectivity during sleep. *Science* 2005;**309**:2228–32.

Mather JA. Cephalopod consciousness: behavioural evidence. *Conscious Cogn* 2008;**17**:37–48.

Mediano P, Seth A, Barrett A. Measuring integrated information: comparison of candidate measures in theory and simulation. *Entropy* 2019;**21**:17.

Napolitani M, Bodart O, Canali P, *et al*. Transcranial magnetic stimulation combined with high-density EEG in altered states of consciousness. *Brain Injury* 2014;**28**:1180–9.

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;19:979–96.

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. *The Logic of Scientific Discovery*. New York: Harper & Row, 1968.

Putnam H. Minds and machines. In *Dimensions of Mind*, ed. S. Hook, New York: New York University Press, 1960, pp. 57–80.

Radnitzky G. Review: Refined falsificationism meets the challenge from the relativist philosophy of science. *Br J Philos Sci* 1991;**42**:273–284.

Reardon S. Rival theories face off over brain's source of consciousness. Science 2019;**366**:293–293.

Richards BA, Lillicrap TP, Beaudoin P, *et al*. A deep learning framework for neuroscience. *Nat Neurosci* 2019;**22**:1761–70.

Rosenthal DM. How many kinds of consciousness? *Conscious Cogn* 2002;**11**:653–65.

Rounis E, Maniscalco B, Rothwell JC, *et al*. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* 2010;**1**:165–75.

Schäfer AM, Zimmermann HG. Recurrent neural networks are universal approximators. In: *International journal of neural systems*. Springer, 2007;**17**:253–63.

Sergent C, Dehaene S. Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *J Physiol Paris* 2004;**98**:374–84.

Seth AK. Models of consciousness. *Scholarpedia* 2007;**2**:1328.

Seth AK. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn Neurosci* 2014;**5**:97–118.

Skinner BF. The behavior of organisms: an experimental analysis. *Appleton-Century*, Cambridge, Massachusetts: B.F. Skinner Foundation. 1938.

Solomonoff RJ. A formal theory of inductive inference. Part I. *Inform Control* 1964;**7**:1–22.

Tononi G. An information integration theory of consciousness. *BMC Neurosci* 2004;**5**:42.

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

Giulio T. Why Scott should stare at a blank wall and reconsider (or, the conscious grid). In: *Shtetl-Optimized: The Blog of Scott Aaronson*. Available online: http://www. scottaaronson.com/blog, 2014.

Tsuchiya N, Wilke M, Frässle S, *et al*. No-report paradigms: extracting the true neural correlates of consciousness. *Trends Cogn Sci* 2015;**19**:757–70.

Tsuchiya N, Taguchi S, Saigo H. Using category theory to assess the relationship between consciousness and integrated information theory. *Neurosci Res* 2016;**107**:1–7.

Tsuchiya N, Andrillon T, Haun A. A reply to "the unfolding argument": beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. PsyArXiv, 2019.

Turing AM. Computing machinery and intelligence. Mind 1950;**59**:433–60.

Turing AM. On computable numbers, with an application to the entscheidungsproblem. *Proc Lond Math Soc* 1937;**2**:230–65.

Wenzel M, Han S, Smith EH, *et al*. Reduced repertoire of cortical microstates and neuronal ensembles in medically induced loss of consciousness. *Cell Syst* 2019;**8**:467–74.

Wolfram S. Cellular automata as models of complexity. *Nature* 1984;**311**:419.

Wu T, Tegmark M. Toward an artificial intelligence physicist for unsupervised learning. *Phys Rev E* 2019;**100**:033311.