Falsification and consciousness

Johannes Kleiner\textsuperscript{1} and Erik Hoel\textsuperscript{2}

\textsuperscript{1}Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich
\textsuperscript{2}Allen Discovery Center, Tufts University, Medford, MA, USA

April 8, 2020

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.

1 Introduction

All successful scientific fields grow in complexity over time, moving 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 that have been advanced by different authors, some quite formal and complicated (Koch et al., 2016).

At this point in the field’s development falsification has become an issue. In general, scientific theories should strive to be falsifiable (Popper, 1959). Yet in some fields, particularly theoretical physics, issues of falsifiability have become a point of major contention as some proposed and even popular physical theories may be unfalsifiable (Ellis and Silk, 2014; Woit, 2006). In the search for a scientific theory of consciousness, falsifiability must be considered explicitly to avoid a similar circumstance. This is especially important because consciousness research is unlike other fields in that it is commonly assumed that consciousness itself cannot be directly observed, instead it can only be inferred based off of brain states, which are physical phenomena.

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 properties of 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 widely across different brain regions (Baars, 1997). Specifically, a global neuronal workspace theory has been proposed in which consciousness is the result of an “avalanche” or “ignition” of widespread

\textsuperscript{*}Johannes.Kleiner@lrz.uni-muenchen.de
\textsuperscript{†}erik.hoel@tufts.edu
neural activity created by an interconnected but dispersed network of neurons with long-range connections (Sergent and Dehaene, 2004).

Another avenue of research has been associating the complexity of brain activity with consciousness (Tononi and Edelman, 1998) which evolved into its most specific instantiation: Integrated Information Theory (Tononi, 2004, 2008; Oizumi et al., 2014). Historically, Integrated Information Theory 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 are 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 $\Phi$).

Both Global Neuronal Workspace (GNW) and Integrated Information Theory (IIT) have gained widespread popularity, sparked general interest in consciousness, and lead to dozens if not hundreds of new empirical studies (Massimini et al., 2005; Del Cul et al., 2007; Dehaene and Changeux, 2011; Gossyries 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 that is pre-registering predictions from the two theories so that testing can be conducted in controlled circumstances by researchers across the world (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. While there are too many other relevant theories, e.g., (Lau and Rosenthal, 2011; Chang et al., 2019), to fully discuss, our results and reasoning apply to most if not all contemporary theories, particularly in their ideal and precise 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 non-formalized theories (a positive attribute). Recently one criticism levied against IIT was based on the fact that the theory predicts feedforward neural networks have zero $\Phi$ and recurrent neural networks have non-zero $\Phi$. Since it is known that a given recurrent neural network can be “unfolded” into a feedforward one while preserving its output function, this is argued to render IIT wrong or outside the realm of science (Doerig et al., 2019). Replies have criticised the assumptions which underlie the derivation of this argument (Kleiner, 2019; Tsuchiya et al., 2019).

Here we frame and expand concerns around testing and falsification of theories by examining a much 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 background 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 (based usually on report or behavior) and the contents of consciousness as predicted by the theory given 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. Instead, it requires some comparison, that is, an inference of an agent’s conscious experience, e.g., using report or behavior. For instance, imagine a theory that predicts (based on such internal observables as 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 needs to be based on finding such mismatches between reported experience and prediction.

In the following work, we formalize this by describing the ideal experimental setup for testing a theory of consciousness based on a physical system, its observables, predictions of conscious content from those observables, and inferences by the experimenter based on report or behavior.

We come to a surprising conclusion: it is likely that all contemporary theories of consciousness are already falsified. This is because contemporary testing rests on an assumption of independence wherein an experimenter’s inferences about consciousness are independent from a theory’s predictions, meaning that inferences can be kept steady while predictions vary. To demonstrate the problems this independence creates for most 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, and constitutes both a generalization and correction of the unfolding argument 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), or replacing a given neural network with a different recurrent neural network or a feedforward neural network, since both types of networks can approximate any given function (Hornik et al., 1989; Schäfer and Zimmermann, 2006), or any of these with some schema for universal intelligence, like the AIXI model, which is based on Solomonoff induction (Hutter, 2004).

While for some particular theory of consciousness there may be a system where there is no significant mismatch between the theory’s predictions based on internal observables and that system’s reports, if inferences and predictions are independent, there exists some different system with different internal observables but equivalent inferences to the original system. This second system can be “substituted” in for the original one in the testing of the theory. Since internal observables following this substitution are then by definition changed, this will lead to differing predictions of consciousness but with the same inferences by the experimenter about consciousness. In turn this triggers a mismatch between predictions and inferences, thus falsifying 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. We identify the root cause of this issue as the assumption of independence between inferences and predictions. 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 reliance on using 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, e.g., the case in behaviorist theories of consciousness, but arguably also in in Global Workspace Theory (Baars, 2005), the “attention schema” theory of consciousness (Graziano and Webb, 2015) or “fame in the brain” (Dennett, 1993) proposals. We thus also subject this response to a detailed formal analysis. We find that making a theory’s predictions and an experimenter’s inferences dependent leads to unfalsifiability. Correspondingly, this assumption is not compatible with a scientific approach to consciousness either.

Our results show that if independence of prediction and inference holds true, as in cases where report about experiences is relied upon, it is likely that no contemporary theory of consciousness is correct. Each is “pre-falsified,” so to speak. Alternatively, if the reasonable assumption of independence is rejected, theories rapidly become unfalsifiable. While this twin dilemma may seem a highly negative set of conclusions, the positive side of our work is that it constrains the future set of viable theories of consciousness in a useful way. We discuss various paths forward for researchers who seek to develop and test such theories.

2 Formal description of testing theories

Here we provide a formal framework for experimentally testing a particular class of theories of consciousness. The class of theories we consider make predictions about the conscious experience of physical systems based on observations or measurements. This class includes many contemporary theories, including leading theories such as Integrated Information Theory (Oizumi et al., 2014), Global Neuronal Workspace Theory (Dehaene and Changeux, 2004), Predictive Processing (when applied to account for conscious experience (Dolega and Dewhurst, 2020; Clark, 2019; Seth, 2014; Hobson et al., 2014; Hohwy, 2012)) or Higher Order Thought Theory, HOT (Rosenthal, 2002). In some cases, contemporary theories in this class may lack the specificity to actually make predictions in their current form. Therefore, the formalisms we introduce may sometimes describe a more advanced or more developed form of a theory, one that can actually be applied to make predictions.

In the following section, we introduce the necessary background and terms to define how to falsify this class of theories: how measurement of a physical system’s observables results in data sets (section 2.1), how a theory makes use of those data sets to offer a predictions about consciousness (section 2.2), how an experimenter makes an inference about a physical system’s experiences (section 2.3), and finally how falsification of a theory occurs when there is a mismatch between a theory’s prediction and an experimenter’s inference about experiences (section 2.4). In section 2.5 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 dependency.

2.1 Experiments

All experimental attempts to either falsify or confirm a theory of consciousness (in the class we consider) begin by examining some particular physical system which has some specific physical configuration, state, or dynamics, \( p \). 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 Global Neuronal Workspace, \( P \) might denote 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 on the system under consideration. For instance, measurements based on neuroimaging tools like fMRI or EEG have to be carried out in order to obtain information about the brain. This information might be used to create detailed data sets such as functional networks, wiring diagrams, models, or transition probability matrices. To formalize this, we denote by \( \mathcal{O} \) all possible data sets that can result from observations of \( P \). Each \( o \in \mathcal{O} \) is one particular data set, and is the result of carrying out some set of measurements on \( p \). We denote the process of measurement of the observables of \( p \) as \( \text{obs}(p) \):

\[
\text{obs}: P \rightarrow \mathcal{O},
\]

where \( \text{obs} \) is a correspondence, which is a “generalized function” that allows for there to be more than one element in the image \( \text{obs}(p) \). \(^1\) A correspondence is necessary because, for a given \( p \), various possible data sets may arise, e.g., due to different measurement techniques such as fMRI vs. EEG, or due to the stochastic behaviour of the system under measurements, 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 data sets which can result from the experiment.

Note that \( \text{obs} \) describes the experiment, the choice of observables, and all conditions during an experiment that generates the data set \( o \) necessary to apply the theory, which may differ from theory to theory. \(^2\) In all realistic cases, the correspondence \( \text{obs} \) is likely quite complicated. It depends on what measurements are performed and might be difficult to state simply. But for our argument, it 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 \( \text{obs} \), but are rather inferred based on some data set \( o \), and \( o \) does not contain the experiences themselves. \(^3\)

In the subsequent sections, we explore how the data sets in \( \mathcal{O} \) are used to make predictions about the experience of a physical system.

2.2 Predictions

A theory of consciousness makes predictions about the experience of some physical system in some configuration, state, or dynamics, \( p \), based on some data set \( o \). 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 content of conscious experience in contemporary IIT, \(^1\)In the case of a function \( f : A \rightarrow B \), for every \( a \in A \) there is exactly one element \( f(a) \in B \). A correspondence \( g : A \rightarrow B \) can specify more than one element for every \( a \in A \): The image \( g(a) \) is a subset of \( B \). Functions are special cases of correspondences.

\(^2\)For example, in cases where a theory necessitates perturbations, knock-outs, or noise injections, such as IIT, this is included in \( \text{obs} \), since such interventions always manipulate observables then measure the consequence on other observables. That is, \( \text{obs} \) describes the whole experimental setup.

\(^3\)Even though phenomenology, i.e., the structure of experience, is used to derive the axioms of IIT, the applications of the postulates never include experience themselves, rather, calculating the integrated information is done using some third-person data set \( o \).
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 physical system. $\mathcal{O}$ then denotes all data sets that can arise from observations or measurements on $P$. Measuring the observables of $p$ maps to data sets $o \in \mathcal{O}$, which is denoted by the $\text{obs}$ correspondence. $E$ represents the mathematical description of experience given the theory or model of consciousness under consideration. In the simplest case, this is just a set whose element indicate whether a stimulus has been perceived consciously or not, but far more complicated structures can arise (e.g., in IIT). The correspondence $\text{pred}$ describes the process of prediction as a map from $\mathcal{O}$ to $E$.

or the description only of whether a presented stimuli is experienced in GNW or HOT. We sometimes refer to elements $e$ of $E$ as states of experience or simply experiences.

Formally, this means that a prediction considers an experimental dataset $o \in \mathcal{O}$ (determined by $\text{obs}$) and specifies an element of the experience space $E$. We denote this as $\text{pred}$, for “prediction,” which is a map from $\mathcal{O}$ to $E$. The details of how individual data sets 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 $\text{pred}$. However, we have to take into account that a single data set $o \in \mathcal{O}$ may not predict only one single experience. In general, $\text{pred}$ 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 $\text{pred}(o)$. Thus, $\text{pred}$ is also a correspondence 

\[ \text{pred} : \mathcal{O} \rightarrow E. \]

Shown in Figure 1 are the full set of terms needed to formally define how most contemporary theories of consciousness make predictions about experience. So far, what we have said is very general. Indeed, the force and generalizability of our argument comes from the fact that we do not have to define $\text{pred}$ 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 $\Phi$ (Aaronson, 2014; Tononi, 2014). However, debates about counter-intuitive predictions of such a theory are not meaningful by themselves, since $\text{pred}$ 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.

### 2.3 Inferences

As discussed in the previous section, a theory is unfalsifiable given just predictions alone. Falsifications require that the results of $\text{pred}$ 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 that we focus on here assume 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 based on reports about seen vs. unseen images in masking paradigms (Alais et al., 2010). In such common cases the report would be some key press indicating a certain visual experience.

It has been pointed out that such a 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 such 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
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 data set $o \in O$ collected in an experimental trial to some subset of experiences described by the model, $E$.

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 non-humans incapable of verbal report, “report” can be broadly construed as behavior or output.

Despite the number of ways of inferring experiences, all these cases can be broadly described as being a case of inference off of some data. This data might be actual verbal 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 of 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 data set $o$ generated by observation or measurement of the physical system, and the set of postulated experiences in the model of consciousness, $E$:

$$inf : O \rightarrow E.$$  

Defining $inf$ as a function means that we assume that for every experimental data set $o$, 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 affect our results. 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 Figure 2.

It is worth noting that we have used here the same class $O$ as in the definition of the prediction mapping $pred$ above. This makes sense because the inference process also uses data obtained in experimental trial, such as reports by a subject. Note that this fully compatible with use of data that has been collected in advance. For example, in cases where inference does not actually draw from the observables during the experiment (like an assumption of compos mentis), it must draw from some other observables (how compos mentis was decided), which can be formally included in our setup if necessary. So both $pred$ and $inf$ can be described to operate on the same total data set measured, even though they usually use different parts of this data set (cf. below).

2.4 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 $pred$ just introduced.

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

$$inf(o) \notin pred(o).$$  

This definition can be spelled out in terms of individual components of $E$. To this end, for any given data set $o \in O$, let $e_r := inf(o)$ denote the experience that is being inferred, and let $e_p \in obs(o)$ be one of

---

4 This assumption is not a restriction of our formalism, because if two correspondences to the same space are given, one can turn of of them into a function by coarse-graining. This makes the definitions involving experimental falsification, such as Definition 2, conceptually and mathematically simple.
the experiences that is predicted based off of some data set. Then (2) simply states that we have \( e_p \neq e_r \) for all possible predictions \( e_p \). None of the predicted states of experience is, in other words, equal to the inferred experience.

What does Equation (2) mean? There are two cases which are possible. Either the prediction based on the model of consciousness is correct, so that according to this equation the inferred experience is wrong and therefore the particular prediction at \( o \) cannot be tested. Alternatively, the prediction could be wrong, so that in this case the model is falsified. In short: Either the prediction process or the inference process is wrong.

We remark that if there is a data set \( o \) on which the inference procedure \( \text{inf} \) or the prediction procedure \( \text{pred} \) 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 data sets \( o \) for which both procedures are defined, i.e. for which both a prediction and a falsification is possible.

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, \( \text{inf} \) and \( \text{obs} \) will make use of different part of the data obtained in an experimental trial. E.g., in the context of IIT or GNW, data about the internal structure and state of the brain will be used for the prediction. This 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 Figure 3. We use the following notation:

- \( o_i \) For a chosen data set \( o \in \mathcal{O} \), we denote the part of the data set 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 data set \( o \in \mathcal{O} \), we denote the part of the data set which is used for inferring the state of experience by \( o_r \) (adding the subscript \( r \), similar to how restrictions are denoted in sets). We call it the inference data in \( o \).

2.5 Summary

In summary, the formalism we have introduced so far describes an experimental test of a theory of consciousness. We have introduced the following notion:
$P$ denotes a class of physical systems, 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 (data sets) 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.

$O$ is the class of all possible data sets that can result from observations or measurements of the systems in the class $P$. Any single experimental trail results in a single data set $o \in O$, whose data is 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 data set $o$. It is therefore a mapping from $O$ 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 $e_r$.

### 3 The substitution argument

In this section, we subject the notion of falsification as used in contemporary experiments to a thorough formal analysis. We do this by considering changes of physical systems we call substitutions.

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. Note that a specific case of such a substitution, the unfolding of a reentrant neural network to a feed-forward 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 a 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 Section 4.

We begin by defining what a substitution is in Section 3.1, show that it implies falsifications in Section 3.2, and analyze the particularly problematic case of universal substitutions in Section 3.3. In Section 3.4, we prove that universal substitutions exist if prediction and inference data is independent and give some examples of already-known cases.

#### 3.1 Substitutions

Substitutions are changes of physical systems that leave the inference data invariant. In order to define formally what a substitution is, we work with the inference content $o_r$ of a data set $o$ as introduced in Section 2.4. 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 correlation $obs$ which describes the relation between physical systems and measurement results, this can be defined as

$$ P_{o_r} := \{ p \in P \mid o_r \in obs(p) \} , $$

where $obs(p)$ denotes all possible data sets that can be measured if the system $p$ is under investigation and where $o_r \in obs(p)$ is a shorthand for $o \in obs(p)$ 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.
Figure 4: This picture illustrates substitutions. Assume that some data set \( o \) 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 data sets is different. Different in fact to such an extend that the predictions of consciousness based on these datasets are incompatible (illustrated by the non-overlapping circles on the right). Here we have used that by definition of \( P_{o_r} \) every \( p \in P_{o_r} \) yields at least one data set \( o' \) with the same inference content as \( o \) and have identified \( o \) and \( o' \) in the drawing.

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

\[
\text{pred} \circ \text{obs}(p) \cap \text{pred} \circ \text{obs}(S(p)) = \emptyset. \tag{4}
\]

In words, this definition 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 \( a_r \), 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 Figure 4. We remark that if \( \text{obs} \) were a function, so that \( \text{obs}(p) \) only contained one element, Equation (4) would be equivalent to \( \text{pred}(\text{obs}(p)) \neq \text{pred}(\text{obs}(S(p))) \).

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 call this case a universal substitution:

**Definition 3.2.** There is a universal substitution if there is an \( o_r \)-substitution for every \( o_r \). \(^6\)

### 3.2 Substitutions imply falsifications

The force of our argument comes from the fact that if there are substitutions (transformation \( S \) which satisfy (4)), then this necessarily leads to mismatches between inferences and predictions. This is shown by the following lemma.

**Lemma 3.3.** If there is an \( o_r \)-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 \( o \in \text{obs}(p) \) 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

\[
\text{pred}(o) \cap \text{pred}(o') = \emptyset. \tag{5}
\]

\(^5\)Here, the \( \circ \) indicates the composition of the correspondences \( \text{obs} \) and \( \text{pred} \) to give a correspondence from \( P \) to \( E \), which could also be denoted as \( \text{pred}(\text{obs}(p)) \), and \( \cap \) denotes the intersection of sets.

\(^6\) We recall that according to the notation introduced in Section 2.4, the inference content of any data set \( o \in \mathcal{O} \) is denoted by \( o_r \) (adding the subscript \( r \)). Thus the requirement is that there is an \( o_r \)-substitution for every inference data that can pertain in the experiment under consideration (for every inference data that is listed in \( \mathcal{O} \)).
Since, however, \( o_r = o_r' \), we have \( \inf(o) = \inf(o') \). Thus we have either \( \inf(o) \notin \text{pred}(o) \) or \( \inf(o') \notin \text{pred}(o') \), or both. Thus there is either a falsification at \( o \), a falsification at \( o' \), or both. \( \square \)

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.

### 3.3 Universal substitutions imply complete falsification

In Section 2.4, we have defined falsifications for individual data sets \( o \in \mathcal{O} \). Using the ‘insight view’ of single data sets, 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 data sets 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 data sets \( o \) and \( o' \) used in that proof both have inference content \( o_r \). \( \square \)

This finally allows us to show our main result. It shows that if a universal substitution exists, the model of consciousness under consideration indeed is highly problematic. 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. \( \square \)

What is the meaning of Proposition 3.6? 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 Section 2.4), there is a data set \( o \) which contains the report’s data and for which we have

\[
\inf(o_r) \notin \text{pred}(o),
\]

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 model 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 rubbish. 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.

In Section 4, we interpret this result. But first, we discuss under which circumstances a universal substitution exists.

### 3.4 When does a universal substitution exist?

In the last section, we have seen that if a universal substitution exists, this has strong consequences for a theory under consideration. In this section, we discuss whether a universal substitution exists.
3.4.1 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 data sets. This is the reason why \( \text{pred} \) is a correspondence, the most general object one could consider. For the purpose of this section, we add a gentle assumption which restricts \( \text{pred} \) 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 don’t consider theories of consciousness which have only a single prediction.

A simple way of formalizing this assumption would be to assume that for every \( o \in \mathcal{O} \), there is a \( o' \in \mathcal{O} \) with \( \text{pred}(o) \cap \text{pred}(o') = \emptyset \). This ensures that \( \text{pred} \) can make multiple predictions: that is, that for every \( \text{pred}(o) \) there is some \( \text{pred}(o') \) that gives a different prediction. Indeed, if \( \text{pred}(o) \) were all of \( E \), that would be a meaningless prediction.

However, this formalization is too naive, as it ignores the fact that data sets come from physical systems \( p \) when measured in experimental trials. 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 data sets which can be generated by physical systems \( p \) that also generate \( o \). When applying our previous definitions, this can be fleshed out as

\[
\bar{o} = \{ o' \mid \exists p \text{ such that } o \in \text{obs}(p) \text{ and } o' \in \text{obs}(p) \}.
\]

Using this, we can state our minimal information assumption in a way that is compatible with the general setup displayed in Figure 2:

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

\[
\text{pred}(\bar{o}) \cap \text{pred}(\bar{o}') = \emptyset.
\]

3.4.2 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 data set. What is more, they are independent, in the sense 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

\[
\nu : P \rightarrow P
\]

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 \( p \in P \).

Note that if inference and prediction data are not independent, e.g. because they have a common cause, problems of tautologies loom large, cf. Section 4.

3.4.3 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 \( o \in \mathcal{O} \), a \( o_r \)-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 \( \nu : P \rightarrow P \) such that \( o_i \in \text{obs}(p) \), \( o'_i \in \text{obs}(\nu(p)) \), \( o_r \in \text{obs}(p) \) and \( o_r \in \text{obs}(\nu(p)) \). By Definition (7), the first two of these four

\[
\text{pred}(\bar{o}) \cap \text{pred}(\bar{o}') = \emptyset.
\]

\[
\text{pred}(\bar{o}) \cap \text{pred}(\bar{o}') = \emptyset.
\]

\[
\text{pred}(\bar{o}) \cap \text{pred}(\bar{o}') = \emptyset.
\]
conditions imply that $\text{obs}(p) \subset \bar{o}$ and $\text{obs}(\nu(p)) \subset \bar{o}'$. Thus Equation (8) applies and allows us to conclude that

$$\text{pred}(\text{obs}(p)) \cap \text{pred}(\text{obs}(\nu(p))) = \emptyset.$$  

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

$$S : P_o \rightarrow P_o',$$

which in light of the first part of this proof exhibits at least one $p \in P_o$, 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 data set $o$, there is another data set $o'$ which makes a non-overlapping 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 Proposition 3.9 and Proposition 3.7. □

This concludes the statement of our main claim. In the next section, we give several examples of universal substitutions, before discussing various possible responses to our result in Section 4.

### 3.4.4 Examples of report and prediction 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 inferences that rely on report and prediction data. We take report to mean output, behavior, or verbal report itself.

**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 non-injective, i.e. some nodes are not part of the output. E.g., 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 single-layer ANNs and recurrent ANNs can approximate any given function (Hornik et al., 1989; Schäfer and Zimmermann, 2006). Since reports are just some function, there are viable universal substitutions that provably exist.

A special case thereof is the unfolding transformation considered in Doerig et al. (2019) in the context of IIT. The arguments in this paper 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 results (using a different characterization of theories of consciousness than minimally informative) has been shown in Kleiner (2019).

**Universal computers.** Turing machines are extremely different 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, 2009). Therefore, if one believes it is possible for a Turing machine (or some sped-up variant thereof) to pass the Turing test, one should accept the substitution argument. 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, 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 via an algorithm 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 substitution of reports. 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 (compared to, say, the human brain).

These are all real and viable classes systems that are used everyday in science and engineering which all provide different viable universal substitutions if inferences are based on reports or outputs. We have avoided possible but less reasonable examples of universal substitutions, like astronomically-large look-up ledgers of reports.

The above examples 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. So 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 (since inferences are generally based solely on reports).

As an example of our Main Theorem 3.10, we consider the case of IIT. For this theory it has already been directly proven, with multiple examples, that inferences and predictions are independent. Since the theory is normally applied in Boolean networks, logic gates, or artificial neural networks, we take report to mean “output.” It has already been proven that systems with different observables and corresponding different predictions from IIT, can have identical input/output (and therefore reports or inferences about report) (Albantakis and Tononi, 2019). Similar research has shown that \( \Phi \) is not invariant under isomorphisms of labeling schemes, despite those schemes having identical input/output (Hanson and Walker, 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. Note that this has been shown even in the strictest sense of universality wherein the reports of the system are not only highly similar, but completely identical (even the dynamics of the internal observables, but not their causal relationships, are identical in some already demonstrated cases).

One immediate response may be to argue that universal substitution, in the strictest sense of universality (for each and every possible inference content), is impossible despite the examples we have shown. That is, even a computer capable of passing the Turing test might not have, for instance, a sense of smell, and therefore would not count as truly universal. Alternatively, it might be argued that perhaps there are some classes of problems that a Turing machine cannot solve but the brain can (Penrose, 1994).

However, the issues from independence do not actually require universal substitutions. Instead, it requires only that there be independence between \( o_i \) and \( o_i' \). So even if \( o \) and \( o' \) do not fully overlap there can still be mismatch, and therefore falsification, which can be created via partial substitutions. Therefore, there is strong evidence to believe that data independence given the normal assumptions behind testing scientific theories of consciousness while relying on report, behavior, or output.

3.4.5 What’s beyond reports?

It may be argued that the examples we have given here do show that data independence holds, but only when \( inf \) is based on reports (or outputs or behavior more generally). This is true. Yet report being the basis of inferences about consciousness is an assumption shared widely, perhaps universally, within contemporary consciousness research.

The reasons for this is an attempt to be as agnostic as possible, since it is unknown a priori (without a theory) which systems are conscious, without looking at behavior, output, or report. In a maximally agnostic view, \( inf \) should solely depend on the richness of the available report and/or the intelligence of the system under question. That is, an experimenter’s confidence in inferences about experiences should grow with richness of report, intelligence of behavior, or some combination of the two. This is quite reasonable. Yet relying on solely on reports entails data independence and therefore creates fragility in theories and allows...
for falsifications via substitutions.

One response to the availability of substitutions might be to restrict all testing of theories of consciousness solely to humans. If testers simply refuse to accept reports about experiences from non-humans, testing can proceed apace. There are some significant downsides to this strategy, since this is equivalent to saying that all inferences in non-humans are wrong. First, many theories of consciousness clearly wish to answer questions about non-human organisms and things like artificial intelligence. The consequence of such a restriction is that such predictions about consciousness outside of humans are no longer falsifiable. This drastically cuts down the reach of theories of consciousness. Second, and more fundamentally, there are likely variations (as in 3.8) available within the brain wherein \( o_r \) is fixed but \( o_i \) varies, meaning that there is still data independence even just in brains alone.

Another reasonable response to these issues is to explore setups and theories wherein there is data dependency, such as when predictions are based on reports or alternatively accessibility to report. As we will next show, theories that imply data dependency have their own issues and flaws, most notably that of triviality or unfalsifiability. Therefore, consciousness research is caught between a rock and a hard place. Clearly, however, our results show that inferences based on reports, behavior, or output alone cannot be used to test theories of consciousness.

4 Responses to the substitution argument

In this section, we list several possible responses to the previous results.

4.1 Inference and prediction data are not independent

The simplest response to our result is to claim that inference and prediction data are not independent. To show that this claim has important negative consequences, we consider first the complete opposite case to independence: the case in which inference and prediction data are strictly dependent.

4.1.1 Inference and prediction are strictly dependent

In order to analyse the case where inference and prediction data are dependent, we first need to say something about unfalsifiable theories of consciousness. Clearly, the goal of the scientific study as a whole is to find, eventually, a theory of consciousness that cannot be falsified in experimental situations, because it is compatible with all experimental evidence. This is a necessary condition (though not sufficient) any purportedly “true” theory of consciousness needs to satisfy. Therefore, unfalsifiability per se is not a bad thing. However, unfalsifiability can also be 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 empirically unfalsifiable.

Such empirically unfalsifiable theories can be identified neatly in our formalism. To see how, recall that \( \mathcal{O} \) denotes the class of all data sets that can result from an experiment under investigation. Put differently, it contains all data sets that can appear in nature when probed in the experiment. This is not the class of all possible data sets of type \( \mathcal{O} \) one can think of. Many data sets 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 data sets as:

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

By construction, \( \mathcal{O} \) is a subset of \( \overline{\mathcal{O}} \), which describes which among the possible data sets actually arises across experimental trails. Hence, \( \mathcal{O} \) also determines which theory of consciousness is compatible with (i.e. not falsified by) the experimental investigation.

Introduction of \( \overline{\mathcal{O}} \) allows us to distinguish the pathological and non-pathological cases of unfalsifiability mentioned above. Whereas any purportedly true theory should only be unfalsifiable 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 unfalsifi-ability. 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 \( \mathcal{O} \), 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 \( o \in \mathcal{O} \), we have \( o_i = f(o_r) \).\(^8\)

This definition should be understood as placing a condition on \( \mathcal{O} \): that there is a function \( f \) as described. If inference and prediction are assumed to be dependent, the class of possible data sets is smaller than if there were no dependence. The definition refers to \( \mathcal{O} \) 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.

This definition of strict dependence is satisfied, for example, if inference data is equal to prediction data, i.e. if \( o_i = o_r \), where \( f \) is simply the identity. This is, for example, the case for behaviourist theories (Skinner, 1938) of consciousness, where consciousness is equated directly with report or behavior. It is also the case for early functionalist theories of consciousness based on behavior or input/output (Putnam, 1960).

Another example of strict dependence is the case where prediction data is always a subset of the inference data:

\[
o_i \subseteq o_r.
\]

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, 1993) proposals.

In all these cases, consciousness is generated by – and hence needs to be predicted via – some form of inference on the system about 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 which models access consciousness alone.

Our second main theorem proves this formally.

**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 data sets explicitly. The possible pairings that can occur in an experiment are given by

\[
\mathcal{O}_{\text{exp}} := \{ (o_i, o_r) \mid o \in \mathcal{O} \},
\]

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 \( \mathcal{O} \), 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

\[
\mathcal{O}_\text{all} := \{ (o_i, o'_r) \mid o \in \mathcal{O}, o' \in \mathcal{O} \}
\]

\[
=( o_i, o'_r) \mid o_i \in \mathcal{O}_i, o'_r \in \mathcal{O}_r \}.
\]

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

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

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

\(^8\) We remark that this formalization of dependency is more fundamental than the one used in Definition 3.8, in that any if inference and prediction data are independent as defined there, they are not dependent as defined here. These slightly different ways of formalizing are necessary to prove our results given the parsimonious formalization we apply throughout the paper. More detailed comparison would require introduction of more formalism, which we would like to avoid. We also remark that both Definition 3.8 and Definition 4.1 could be stated in a more symmetric way. Again, this is unnecessary and would only make matters more complicated. In fact, there exists an inverse substitution argument, which we do not discuss here, with subscripts \( i \) and \( r \) replaced everywhere.
where we have symbols $d_i$ and $d_r$ to denote prediction and inference data independently of any data set $o$.

Assume that the first condition in this expression, $d_r = o_r$ holds for some $o \in \emptyset$. 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 data set $o$ 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

$$\mathcal{O}_{\text{all}} = \{ (d_i, d_r) \mid \exists o \in \emptyset \text{ such that } d_r = o_r \text{ and } d_i = o_i \}.$$  \hspace{1cm} (15)

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

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

The gist of this proof is that if inference and prediction are strictly dependent, then as far as the inference and prediction contents go, $\emptyset$ and $\overline{\emptyset}$ are the same. I.e., the experiment does not add anything to the evaluation of the theory. It is sufficient to know only all possible data sets to decide whether there is a falsification. In practice, 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 ad odds with the role of empirical evidence required in any scientific investigation. Thus such theories are empirically unfalsifiable.

To give an 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). I.e., the prediction gives the same experience as the one inferred from what is accessible to report. Therefore, 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 $o_i \subseteq o_r$. Unless something goes wrong, the participant in the experiment should be able to deterministically report what is accessible to report, by definition of that information as accessible. Since the predicted contents are always precisely what can be reported, there can never be any mismatch between reports and predictions. However, this is not only the case for $\mathcal{O}_{\text{exp}}$ but also, in virtue of the theory’s definition, for all possible data sets, i.e., $\mathcal{O}_{\text{all}}$. Therefore such theories are empirically unfalsifiable in that experiments add no information to whether the theory is true or not, and therefore such theories are empirically uninformative or tautological.

This concludes our treatment of the case where inference and prediction data are strictly dependent. Next, we consider other cases of dependencies.

### 4.1.2 Fine-tuned dependence

When combined, our main theorems show that both independence and strict dependence of inference and prediction data are problematic and thus must not be assumed in an experimental investigation. This raises the question of what happens in the case that inference and prediction are dependent, but not strictly dependent. Is there some middle ground which may work?

This issue is worth future investigation, and a full formal analysis goes beyond the scope of this paper. However, there are some things we can point out which are not encouraging.

First, it is important to note that even if a weaker form of independence holds than the one we have worked with in Definition 3.8, there will still be available substitutions. The difference is only that these might not be universal, i.e., might not imply the worst possible scenario (where one single inference that is to be trusted implies that a theory is wrong). But in general they will still cause problems of theories of consciousness, and if inference procedures have any credibility, even non-universal substitutions will imply the theory under consideration is wrong. We expect that the same holds true for strict dependence, so that theories which imply inference and prediction to be close to strict dependence will have to be dismissed.
The only remaining case that has a chance to avoid these problems is a case where inference is fine-tuned to match the prediction procedure in just the right way so as to avoid falsification but also unfalsifiability.

Indeed, the definition of an \( inf \) that does not falsify a \( pred \) for any data set in \( \mathcal{O}_{\text{exp}} \) but does falsify a \( pred \) for some data sets in \( \mathcal{O}_{\text{all}} \) would render a theory of consciousness “true” under our formalisms. It may be possible to construct such a theory. However, there are already notable problems, for this seems to require an \( inf \) that is equally as complex as \( pred \), in that a theorist is constructing \( inf \) as much as \( pred \). It is usually taken to be important that \( inf \) has independent justification from \( pred \), which seems difficult to achieve in the case of fine-tuning of \( inf \). That is, it would no longer be the case that inferences rely simply on verbal report or output, but something much more complex, and therefore difficult to justify. With this independent justification missing, the complexity of a fine-tuned \( inf \) implies that there is no longer any bias as to whether \( pred \) or \( inf \) is wrong during a mismatch. Thus even if a falsification occurs, it may be easy to blame this on \( inf \) and proceed with a modified version thereof. Indeed, if we allow for any \( inf \) to be constructed of arbitrary complexity, then it is likely that for any \( pred \) an \( inf \) can be constructed that makes those predictions “true.”

Thus, while according to our present results, there does seem to be room for saving the usual falsifications scheme, it is clear that to do so, inferences about consciousness must be complexified far beyond report, behavior, or output. This may prove its own challenge as theories of inference must be equally as complex as theories of consciousness themselves.

4.2 Theories could be based solely on phenomenology

One response to the issue of data independence/dependence identified here is to accept that a theory of consciousness may be unfalsifiable but can be judged by other characteristics. This is similar to arguments around certain physical theories, like String Theory. Some have argued that falsifiability is not necessary for some physical theories but that they can be judged by elegance or parsimony instead (Carroll, 2018). Others have argued that this makes such theories beyond science (Ellis and Silk, 2014; Woit, 2006).

In addition to elegance and parsimony, in consciousness science, one could also consider a theory’s fit with the phenomenology of experience. Examples that pursue this fit with phenomenology are IIT (Tononi, 2008) and the isomorphism view (Tsuchiya et al., 2019). The underlying idea here is that the development of a theory of consciousness should aim to establish an isomorphism between phenomenal consciousness and physical substrate (in IIT, this takes the form of axioms about phenomenology and their translation to mathematical postulates).

Any proposal to negate falsifications in the scientific study of consciousness imply a substantial change of scientific methodology, moving away the ideas that have enabled much progress in other disciplines as they became more advanced. Additionally, such proposals face a fundamental inverse problem when dealing only with parsimony/elegance and fit with phenomenology, as alone the degree of introspective detail we have about our own phenomenology may not provide enough information to fit a unique \( pred \) function.

4.3 Theories could solve the Hard Problem

A theory of consciousness might not need to be falsifiable in the sense we have laid out here if it solves the Hard Problem of Consciousness (Chalmers, 1996), i.e. if it provides an a priori logical reduction of conscious experience to the physical domain. This would take the form of a \( pred \) where it is necessary that a physical system in a certain configuration, state, or trajectory, is associated with certain contents consciousness, and there is no way such a thing could not be associated with these contents. In such a case, any \( inf \) function could be considered useless, since even in the cases of mismatch there would be no doubt to believe \( pred \) over \( inf \), due to the necessity of the predictions.

4.4 Experience is an observable during testing

One easy way to avoid our conclusions is to claim that experience is an observable when testing theories of consciousness. In terms of our notation (Figure 2), this would mean that \( \text{obs} \) would include measuring and identifying experiences themselves, and therefore be a map from \( P \) and \( E \) to \( \mathcal{O} \). Equivalently this amounts
to assuming that in an experiment, the “actual” experience $e$ of a system is accessible, next to $e_r$ and $e_p$. This is a case to which our results do not apply.

While promising at first sight, deeper inspection of this idea raises several severe issues. Most notably, it is unclear how experience could be represented in experimental data. Science is an intersubjective endeavour (Nagel, 1989), and so in virtue of data being of use for science, it seems to always be associated with a third-person perspective, and hence cannot “contain” experience.

Another option compatible with this line of response is to consider “direct observation” of experience, meaning that predictions should refer to the experience of the experimenter. Predictions of this sort would constitute a significant change to scientific methodology, whose credibility needs to be evaluated in detail.

More reasonably, consciousness may be indirectly observable in an experiment. This would require that the presence or absence of experience itself changes the configuration, trajectory, or states of physical systems, in other words that the physical is not (causally) closed (Kim, 1998). If a theory of consciousness does not describe the physical as closed, a whole range of predictions are possible: predictions which concern the observable, 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.

5 Conclusion

In this paper, we have subjected the usual scheme for testing theories of consciousness to a thorough formal analysis. We have shown that deep problems are inherent in this scheme which need to be addressed before designing any experimental investigation.

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 number of assumptions that are usually being made, when taken together, are responsible for this problem. Most notably, an experimenter’s inference about consciousness and a theory’s predictions are generally implicitly assumed to be independent during testing. This is because normally the science of consciousness relies on reports, behavior, or output of systems to provide inferences about consciousness, which however implies independence between inferences and predictions. As we formally prove, if this independence holds and reports are relied upon, substitutions or changes to physical systems are possible that falsify any given contemporary theory, under minimum assumptions. However, if an experimenter’s inferences about consciousness and a theory’s predictions are instead considered to be strictly dependent, unfalsifiability follows. Theories of consciousness seem caught between Scylla and Charybdis.

There may be several paths forward that avoid these dilemmas. We have discussed the various possible responses to our results in Section 4. The dilemmas outlined here usefully constrain the set of possible testable theories of consciousness, helping to draw the negative space around a viable scientific theory.

6 Acknowledgments

Funding: This research was supported by the Allen Discovery Center program through The Paul G. Allen Frontiers Group (12171). Author contributions: J.K and E.H. conceived the project and wrote the article. Competing interests: The authors declare no competing interests.

References

Scott Aaronson. Why i am not an integrated information theorist (or, the unconscious expander). Shtetl-Optimized: The Blog of Scott Aaronson, 2014.

David Alais, John Cass, Robert P’Shea, and Randolph Blake. Visual sensitivity underlying changes in visual consciousness. Current biology, 20(15):1362–1367, 2010.

Larissa Albantakis and Giulio Tononi. Causal composition: Structural differences among dynamically equivalent systems. Entropy, 21(10):989, 2019.
Bernard J Baars. In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.

Bernard J Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, 150:45–53, 2005.

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

Randolph Blake, Jan Brascamp, and David J Heeger. Can binocular rivalry reveal neural correlates of consciousness? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1641):20130211, 2014.

Ned Block. How can we find the neural correlate of consciousness? *Trends in neurosciences*, 19(11):456–459, 1996.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Berndrath Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Sean M Carroll. Beyond falsifiability: Normal science in a multiverse. *Why Trust a Theory?*, page 300, 2018.

Michael A Cerullo. The problem with phi: a critique of integrated information theory. *PLoS computational biology*, 11(9):e1004286, 2015.

David J Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford university press, 1996.

Acer YC Chang, Martin Biehl, Yen Yu, and Ryota Kanai. Information closure theory of consciousness. *arXiv preprint arXiv:1909.13045*, 2019.

Andy Clark. Consciousness as generative entanglement. *The Journal of Philosophy*, 116(12):645–662, 2019.

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

Francis Crick and Christof Koch. Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences*, volume 2, pages 263–275. Saunders Scientific Publications, 1990.

Francis C Crick and Christof Koch. What is the function of the claustrum? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1458):1271–1279, 2005.

Stanislas Dehaene and Jean-Pierre Changeux. Neural mechanisms for access to consciousness. *The cognitive neurosciences*, 3:1145–58, 2004.

Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

Antoine Del Cul, Sylvain Baillet, and Stanislas Dehaene. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS biology*, 5(10):e260, 2007.

Daniel C Dennett. *Consciousness explained*. Penguin uk, 1993.

Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H Herzog. The unfolding argument: Why iit and other causal structure theories cannot explain consciousness. *Consciousness and cognition*, 72:49–59, 2019.

Krzysztof Dolega and Joe E Dewhurst. Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, pages 1–26, 2020.

George Ellis and Joe Silk. Scientific method: Defend the integrity of physics. *Nature News*, 516(7531):321, 2014.
Chris Frith, Richard Perry, and Erik Lumer. The neural correlates of conscious experience: An experimental framework. *Trends in cognitive sciences*, 3(3):105–114, 1999.

Rick O Gilmore, Michele T Diaz, Brad A Wyble, and Tal Yarkoni. Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1396(1):5, 2017.

Philip Goff. *Consciousness and fundamental reality*. Oxford University Press, 2017.

Olivia Gosseries, Haibo Di, Steven Laureys, and Mélanie Boly. Measuring consciousness in severely damaged brains. *Annual Review of Neuroscience*, 37:457–478, 2014.

Michael SA Graziano and Taylor W Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in psychology*, 6:500, 2015.

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

J Allan Hobson et al. Consciousness, dreams, and inference: The cartesian theatre revisited. *Journal of Consciousness Studies*, 21(1-2):6–32, 2014.

Jakob Hohwy. Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, 3:96, 2012.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Marcus Hutter. A gentle introduction to the universal algorithmic agent aixi, 2003.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

Jaegwon Kim. *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT press, 1998.

Johannes Kleiner. On empirical well-definedness of models of consciousness. *PsyArXiv*, 2019.

Johannes Kleiner and Sean Tull. The mathematical structure of integrated information theory. *arXiv preprint arXiv:2002.07655*, 2020.

Christof Koch. *The Feeling of Life Itself: Why Consciousness is Widespread But Can’t be Computed*. Mit Press, 2019.

Christof Koch, Marcello Massimini, Melanie Boly, and Giulio Tononi. Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17(5):307, 2016.

Victor AF Lamme. Towards a true neural stance on consciousness. *Trends in cognitive sciences*, 10(11):494–501, 2006.

Hakwan Lau and David Rosenthal. Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, 15(8):365–373, 2011.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Marcello Massimini, Fabio Ferrarelli, Reto Huber, Steve K Esser, Harpreet Singh, and Giulio Tononi. Breakdown of cortical effective connectivity during sleep. *Science*, 309(5744):2228–2232, 2005.

Pedro Mediano, Anil Seth, and Adam Barrett. Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy*, 21(1):17, 2019.

Thomas Nagel. *The view from nowhere*. Oxford University Press, 1989.
Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, 10(5):e1003588, 2014.

Roger Penrose. *Shadows of the Mind*, volume 4. Oxford University Press Oxford, 1994.

Karl Popper. *The logic of scientific discovery*. Routledge, 1959.

Hilary Putnam. Minds and machines. 1960.

Sara Reardon. Rival theories face off over brain’s source of consciousness, 2019.

David M Rosenthal. How many kinds of consciousness? *Consciousness and cognition*, 11(4):653–665, 2002.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, pages 632–640. Springer, 2006.

Claire Sergent and Stanislas Dehaene. Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *Journal of Physiology-Paris*, 98(4-6):374–384, 2004.

Anil Seth. Models of consciousness. *Scholarpedia*, 2(1):1328, 2007.

Anil K Seth. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, 5(2):97–118, 2014.

BF Skinner. The behavior of organisms: an experimental analysis. *Appleton-Century*, 1938.

Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5(1):42, 2004.

Giulio Tononi. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3):216–242, 2008.

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

Giulio Tononi and Gerald M Edelman. Consciousness and complexity. *science*, 282(5395):1846–1851, 1998.

Naotsugu Tsuchiya, Melanie Wilke, Stefan Frässle, and Victor AF Lamme. No-report paradigms: extracting the true neural correlates of consciousness. *Trends in cognitive sciences*, 19(12):757–770, 2015.

Naotsugu Tsuchiya, Thomas Andrillon, and Andrew Haun. A reply to “the unfolding argument”: Beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. *PsyArXiv*, 2019.

Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

Alan Mathison Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society*, 2(1):230–265, 1937.

Michael Wenzel, Shuting Han, Elliot H Smith, Erik Hoel, Bradley Greger, Paul A House, and Rafael Yuste. Reduced repertoire of cortical microstates and neuronal ensembles in medically induced loss of consciousness. *Cell systems*, 8(5):467–474, 2019.

Peter Woit. *Not even wrong: The failure of string theory and the search for unity in physical law*. Basic Books (AZ), 2006.

Stephen Wolfram. Cellular automata as models of complexity. *Nature*, 311(5985):419, 1984.

Tailin Wu and Max Tegmark. Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100(3):033311, 2019.