The Closure of the Physical is Unscientific

Johannes Kleiner[1,2] and Stephan Hartmann[1,3]

[1]Munich Center for Mathematical Philosophy
[2]Munich Graduate School of Systemic Neurosciences
[3]Munich Center for NeuroSciences - Brain and Mind

Abstract We analyze the implications of the closure of the physical for experiments in the scientific study of consciousness when all the details are considered, especially how measurement results relate to physical events. It turns out that the closure of the physical implies that no experiment can distinguish between two theories of consciousness that obey this assumption. Therefore, the closure of the physical is incompatible with scientific practice. This conclusion points to a fundamental flaw in the paradigm underlying most of the experiments conducted to date.

1. Introduction

The closure of the physical is a central assumption in the philosophy of mind and in the scientific study of consciousness [12, 17]. It underlies both functionalist and identity theories of consciousness and is a central component of many, if not all, neuroscientific models of consciousness. However, we will show below that the closure of the physical is untenable in a scientific context because it implies that no experiment can actually distinguish between two theories of consciousness that obey this assumption. It is therefore incompatible with scientific practice and hence *unscientific*.

The central idea of our argument is the observation that in any scientific experiment the measurement results must be stored or transmitted before analysis, and we show that this means that the stored data are determined by the physical properties of a storage device or a transmission channel. In conjunction with the closure of the physical, this means that the stored data are independent of which theory of consciousness is true.

It has already been pointed out that the closure of the physical is a problematic assumption in a scientific context. [18] and [19], for example, make this point with respect to property dualism and qualia epiphenomenalism. Our proof presented below covers the general case. It shows independently of any other metaphysical premises

that one of the central assumptions in the empirical study of consciousness is flawed. This calls into question the theoretical basis of a large number of experiments conducted to date and shows that the hope of basing a functionalist or identity-based understanding of consciousness on empirical observations is null and void.

The remainder of this paper is organized as follows. Section 2 elaborates which theories of consciousness our argument addresses and defines an epistemic version of the closure of the physical. Section 3 identifies a necessary condition for theories of consciousness to be distinguished by empirical data. Sections 4 and 5 discuss the role of empirical data in the scientific study of consciousness and why they supervene on physical events. Section 6 is devoted to the proof of our main claim, and Section 7 shows that the causal closure of the physical, as usually defined ontologically, implies our definition, which ensures that our result holds in full generality. Finally, Section 8 contains some concluding remarks.

2. Theories of consciousness

We use the term *theories of consciousness* to refer to the theories that are tested, compared, or derived in experiments in the scientific study of consciousness, regardless of what metaphysical status of consciousness they presuppose. This includes, for example, Integrated Information Theory [16], Global Neuronal Workspace Theory [14] or Higher Order Thought Theory [1], and in general all scientific theories which adhere to functionalism, identity theory or epiphenomenalism. This also includes illusionist or eliminativist theories that are subject to experimental testing, even though they do not grant consciousness an independent ontological status, but merely aim to explain why someone has the illusion of being conscious [21].

Our results rely on two general facts about theories of consciousness. The first is that theories of consciousness have some commitment with respect to physical events, where *physical events* are the kinds of events that are the subject of natural sciences such as biology, chemistry, neuroscience, and physics. Some theories modify the description of physical events provided by natural science, for example, by postulating changes in the temporal evolution of physical states, as recently in [3], others simply adopt whatever natural science says about physical events without any modification.

The causal closure of the physical is the assumption that for every physical effect, there is a sufficient physical cause. Its key epistemic repercussion (cf. Section 7) is that theories of consciousness must not amend whatever it is that the physical sciences say or imply about physical events. We call this epistemic assumption *closure of the physical*: A theory of consciousness obeys the *closure of the physical* if and only if it does not posit any changes to the physical events explained, predicted or otherwise determined by natural science.

This premise can be expressed concisely in formal terms. To this end, we introduce two sets[1] of event-descriptions. First, for any theory of consciousness T, we denote by $\mathcal{P}_T$ the physical events which T is committed to, for example the firing of some neurons or the instantiation of some functional property. Every element in $\mathcal{P}_T$ is a description of an event that occurs, according to T, in the actual world. The description specifies the event and may include properties or relational information about the event. What exactly a description contains and in which language it is formulated is not of importance here.

Second, we denote by $\mathcal{P}_P$ the physical events which natural science explains, predicts or determines. Whatever it is that natural science says or implies about the physical events in the actual world is part of the class $\mathcal{P}_P$. Each element is in turn a description of an event, including its properties and relations, and we allow that the description is either deterministic or indeterministic.

Since scientific theories are complex, $\mathcal{P}_P$ may not be known or even knowable. And as science progresses over time, $\mathcal{P}_P$ is likely to change over time. For this reason, in what follows, $\mathcal{P}_P$ functions like a variable. It is not important what value this variable actually takes, but only what relationship a theory of consciousness has to this variable.

A theory of consciousness obeys the closure of the physical only if it does not postulate any changes to the class $\mathcal{P}_P$. Thus, it does not replace, change, or add to the description of physical events explained, predicted, or otherwise determined by natural science. This means that for every physical event in $\mathcal{P}_T$ to which a theory of consciousness is committed, there is an element of $\mathcal{P}_P$ that provides a description of that event in one of the languages of a natural science. The descriptions in the two sets may differ in language, but not in content.

In formal terms, this means that there is an *embedding* of $\mathcal{P}_T$ into $\mathcal{P}_P$, i.e. an injective (one-to-one) function ι of the form

$$\iota : \mathcal{P}_T \longrightarrow \mathcal{P}_P \,, \tag{1}$$

which specifies for every physical event and description that the theory of consciousness is committed to the corresponding event and description explained, predicted, or determined by natural science. The existence of this function is the concise meaning of the closure of the physical introduced above: A theory of consciousness T obeys the *closure of the physical* if and only if there exists a function ι as in (1). We will show in Section 7 that the usual reading of the causal closure of the physical implies just that.

3. Experiments

In the scientific study of consciousness, experiments are conducted to falsify, confirm, or distinguish between competing theories of consciousness. The most important

[1]Note that we do not distinguish between classes and sets in this paper.

component of any experiment is measurement, i.e., laboratory operations that produce a set of data which constitutes the result of the measurement.

The second general fact on which our argument is based is that scientific theories of consciousness have something to say about possible measurement results. We assume that any theory allows one to derive, for some experiments and under appropriate auxiliary assumptions, a class of data sets which, according to the theory, may occur as the result of the experiment. This requirement singles out *scientific* theories as those to which our argument applies.

We use the symbol $\mathbb{M}$ to represent an experiment, and furthermore introduce the symbol $\mathcal{O}_{\mathbb{M}}$ to denote all data sets which could result from this experiment according to some assumption or theory. So $\mathcal{O}_{\mathbb{M}}$ denotes the possible measurement results of $\mathbb{M}$ in some context. If an experiment $\mathbb{M}$ only made measurements on one system and everything were deterministic, then there would only be one data set in $\mathcal{O}_{\mathbb{M}}$. But experiments usually consider many systems and things are not deterministic, which is why we have a whole class of data sets that can occur in $\mathbb{M}$.

Given an experiment $\mathbb{M}$ to which a theory T can be applied, we denote the data sets which can occur in $\mathbb{M}$ according T by $\mathcal{O}_T$. In experimental practice, $\mathcal{O}_T$ is deduced from T, making use of approximations and auxiliary assumptions, so that it contains the pre- or retrodictions of T. But in our case we stick to the precise meaning independently of approximations and auxiliary assumptions. Any result $o \in \mathcal{O}_T$ can occur in experiment $\mathbb{M}$ after T, and any $o \notin \mathcal{O}_T$ cannot occur in $\mathbb{M}$ after T. If $o \in \mathcal{O}_T$ occurs, then the probability of T increases (and T is confirmed), and if $o \notin \mathcal{O}_T$ occurs, then the probability of T decreases (and T is disconfirmed). In a Popperian framework, the occurrence of $o \in \mathcal{O}_T$ provides a corraboration of T and the occurrence of $o \notin \mathcal{O}_T$ amounts to a falsification of T.

What matters for our purposes is that if two theories provide the exact same information about which results may or may not occur in an experiment, then these theories cannot be distinguished in that experiment. Theories for which this is the case are empirically indistinguishable. Put concisely in terms of the notation we have just introduced, two theories T and T' are *empirically indistinguishable* if there is no single experiment $\mathbb{M}$ such that $\mathcal{O}_T \neq \mathcal{O}_{T'}$ in $\mathbb{M}$.So if two theories are to be empirically distinguishable, they cannot yield exactly the same class of possible measurement results for each experiment. There must be at least one experiment in which $\mathcal{O}_T \neq \mathcal{O}_{T'}$, so that in this experiment there is at least a chance that a result o occurs which lies in one but not in both classes and is thus consistent with one but not with both theories.[2]

[2]Note that empirical indistinguishability is weaker than empirical equivalence, as defined, for example, in [24] and [25]. Two theories are empirically indistinguishable if they make exactly the same testable statements about experiments to which they are both applicable. Empirical equivalence also requires that the two theories apply to exactly the same experiments.

It is natural to expect that a large number of experiments will not be able to distinguish between two arbitrary theories, since experiments are usually designed with specific theories in mind. Empirical indistinguishability holds only if for two theories there is no experiment at all that can distinguish between them.

If an assumption implies that this is in fact true of *all* theories obeying this assumption, and if there are two or more competing theories which do so, this is obviously problematic. In case such an assumption is implied, all experiments that seek to distinguish between theories become meaningless, and all subsequent differences between theories obeying that assumption untestable. This is incompatible with any empirically based scientific practice, so we consider this a sufficient condition to call such an assumption unscientific. So if an assumption implies that any two different theories obeying that assumption are empirically indistinguishable, we conclude that this assumption is *unscientific*.

We emphasize that this condition is a decidedly weak sufficient condition for a particular assumption not to be scientific. We have by no means proposed a new solution to the notorious demarcation problem. Moreover, the condition is independent of the choice of the preferred account of theory testing. An assumption that is unscientific in this sense undermines any empirical scientific progress in the field in question.

Experiments in the scientific study of consciousness usually use two different types of measurements [2]. First, they make use of what are called *third-person measurements* which employ standard scientific methods. Typical examples are EEG or fMRI recordings. Second, they use what might be called *first-person* or *consciousness-inferring* measurements. This class of measurements has been characterized as using the subject's access to his or her own conscious experience in some way, such as via verbal reports or pressing of a button [15]. More recently, the term *subjective measures of consciousness* has come to refer to these types of measures [10], in contrast to *objective measures* and *no-report paradigms* [23], which infer a subject's state of consciousness indirectly, e.g., by evaluating forced choice tasks [4] or behavioral data such as optokinetic nystagmus and the pupillary reflex [7].

What exactly the difference is between measurements in the first and third person is not important for our purposes. The only important thing is that both types of measurements produce results that need to be analyzed, interpreted or transformed. To do this, they must be stored on a data repository. This fact has implications that we analyze below.

4. Data

We have minimally characterized measurements as laboratory operations that provide a data set that is designated as the result of the experiment. But what does it mean that this data set must be stored on some device? To address this question, let's take a hard disk as an example. A hard disk stores data by magnetizing a thin film

of ferromagnetic material that forms the surface of the hard disk platter. The film is made up of many tiny, sequentially aligned magnetic regions, each of which has a magnetization vector that can point in one of two directions. When data is stored on the disk, the head of the drive arm moves over these areas and changes the magnetization vector by applying electric fields. When reading data from the disk, the actuator arm uses weaker electric fields to sense the magnetization vectors of the areas.

The data stored on the disk is the distribution of magnetization vectors across the magnetic areas in terms of the order of the areas. Two copies of the same disk cannot differ in the data stored on it without differing in at least some magnetization vectors. The data is *determined* by the magnetization vectors.

The crucial thing about the magnetization vectors that determine the data stored on a hard disk is that they are not just properties of the device, but actually *physical properties* of the device, the kind of properties that are the subject of natural science, in this case electromagnetism. Electromagnetism explains their causal properties, such as how the magnetization vector responds to electric fields, and also their dynamic properties, such as how magnetization vectors change over time without interactions.

Accordingly, the occurrence of a particular distribution of magnetization vectors over the ferromagnetic film at a particular time is a *physical event*, the kind of event that is the subject of natural science. It follows that the data stored on the hard disk is determined by a physical event: in this case, the distribution of magnetization vectors over the ferromagnetic film. There is no constraint on why or how this physical event occurs, but once the event occurs, the data stored on the hard disk is determined.

This is true not only for hard drives, but for all data storage devices, such as solid-state drives or flash drives, where the relevant semiconductor properties can only be explained using condensed matter theory and quantum mechanics. But even when data is stored on something as simple as a piece of paper or a spoken word, the data supervene on physical events, namely the distribution of ink molecules on the paper material and air pressure fluctuations, which in these cases represent sound waves.

We can again express this fact succinctly in formal terms. Functions in the mathematical sense of the word are defined to capture exactly those cases where something is completely determined by something else. Let us denote by $\mathbb{P}$ the set or class of all physical events (and descriptions) that can possibly occur in the real world, and by $\mathbb{O}_D$ all records that can possibly be stored on a storage device D. The notion of possibility at issue here is logical possibility. The physical events explained, predicted, or determined by natural science for the actual world form a subset of $\mathbb{P}$, the subset $\mathcal{P}_P$ we introduced above. The same is true for the physical events $\mathcal{P}_T$ to which a theory of consciousness is committed.

The fact that the physical events which occur in the actual world determine the data that is stored on a storage device D can then be represented by a function

$$d_D : P(\mathbb{P}) \longrightarrow P(\mathbb{O}_D)\,, \tag{2}$$

where $P(\mathbb{P})$ is the set of all subsets of $\mathbb{P}$, called the power set of $\mathbb{P}$, and where $P(\mathbb{O}_D)$ is the power set of $\mathbb{O}_D$. The function d_D provides for every logically possible set of physical events $\mathcal{P} \subset \mathbb{P}$ of the actual world a class of data sets $\mathcal{O}_D \subset \mathbb{O}_D$ that could be stored on D at a particular time, so it maps element-wise as

$$d_D : \mathcal{P} \longmapsto \mathcal{O}_D\,. \tag{3}$$

It selects from all physical events which, according to $\mathcal{P}$, are part of the real world those which are relevant for data storage on the device D, e.g. the magnetization vectors in the case of a hard disk. Since $\mathcal{P}$ may contain indeterministic statements, the output of the function may also be indeterministic. For this reason, the output is represented by a class $\mathcal{O}_D$, which may contain more than one record o. However, although $\mathcal{O}_D$ is consistent with indeterminism in physical events, it is completely determined by $\mathcal{P}_P$. This is enforced by the fact that d_D is a function. If D is not instantiated in a set $\mathcal{P}$, the function simply returns the empty set.

In order to use this function in the following, we have to consider two conditions. The first condition arises from the fact that the data stored on a device D corresponding to some physical events is independent of the language used to describe those events. Applied to the embedding ι introduced in (1), this means that

$$d_D\big(\iota(\mathcal{P}_T)\big) = d_D\big(\mathcal{P}_T\big)\,. \tag{4}$$

The content of $\iota(\mathcal{P}_T)$ and $\mathcal{P}_T$ is the same, so also the data stored on D.

The second condition targets situations where one set of physical events completely contains another, e.g. when the latter is a partial description of the former. A set of physical events $\mathcal{P}_2$ completely contains another set $\mathcal{P}_1$ if all event descriptions of $\mathcal{P}_1$ are also contained in $\mathcal{P}_2$, which means that $\mathcal{P}_2$ describes exactly the same events as $\mathcal{P}_1$. It may add to the description of $\mathcal{P}_1$, but it does not change it in any way. Thus, if $\mathcal{P}_1$ includes all the physical events required to instantiate a data repository D, and thus determines the data stored on D, it follows that $\mathcal{P}_2$ also includes these events, so that the data that $\mathcal{P}_1$ and $\mathcal{P}_2$ determine to be stored on D are the same. Whenever we have $\mathcal{P}_1 \subset \mathcal{P}_2$ and D is instantiated in $\mathcal{P}_1$, we have

$$d_D\big(\mathcal{P}_1\big) = d_D\big(\mathcal{P}_2\big)\,. \tag{5}$$

5. Measurement results

We are now ready to apply this result on data storage to experiments in the scientific study of consciousness. The measurements performed in these experiments tend to be quite complex. They may employ advanced brain imaging techniques such as EEG,

ECoG, or fMRI, and require finely tuned equipment and sophisticated analysis to learn about a subject's state of consciousness.

In the case of EEG, ECoG or fMRI recordings, it is relatively clear what the result of such measurements is. It is the data set that the scanner provides after each trial and that is stored in computer memory. In the case of subjective measures, one would normally expect reports or keystrokes to count as results; in the case of objective measures, changes in pupil size and the like. Crucially, however, all of these are physical events. The electrical activity that an EEG electrode measures is as much a physical event as the sound waves that make up a spoken word or the mechanical movements of a button.

Our analysis from the last section allows us to make this point despite the terminological ambiguities about what to count as the result of a measurement. A necessary condition for a record to count as the result of a measurement is that it be stored somewhere. This can be computer memory, but it can also be something simpler like ink on paper or density fluctuations in sound waves. Even data transmission, such as in a cable attached to a button that a person presses, is a form of data storage, albeit of very short duration. So for something to be considered a measurement at all, there must necessarily be a data repository D, so that some of the data stored on D *is* the result of the measurement.

However, we have established above that the data stored on a device D is determined by physical events. Since a part of this data represents the measurement result, the measurement results are also determined by physical events. How these physical events come about – what their causes are – is not constrained by our analysis. The events can have purely physical causes, physical and non-physical causes, or a priori only non-physical causes. Which of these cases applies and with respect to which notion of causality depends on the theory of consciousness.

As before, let us denote by $\mathbb{M}$ an arbitrary but fixed experiment in the scientific study of consciousness, and let us denote by D the data store or stores necessarily used in $\mathbb{M}$ to store the results of the measurement. We have already introduced the symbol $\mathcal{O}_{\mathbb{M}}$ to denote the data sets that, under certain assumptions or theories, could be the possible outcomes of the experiment $\mathbb{M}$. Our analysis from the previous section then shows that $\mathcal{O}_{\mathbb{M}}$ is also determined by the function d_D introduced in (2), namely by restricting d_D to the part of the data stored on D that represents the measurement results. If we denote this restriction by $d_{\mathbb{M}}$ and all data sets that could possibly result from $\mathbb{M}$ by $\mathbb{O}_{\mathbb{M}}$, we obtain a function

$$\begin{aligned} d_{\mathbb{M}} : P(\mathbb{P}) &\longrightarrow P(\mathbb{O}_{\mathbb{M}}) \\ \mathcal{P} &\longmapsto \mathcal{O}_{\mathbb{M}} \,, \end{aligned} \tag{6}$$

which maps any set of physical events $\mathcal{P}$, which could possibly represent the physical events of the actual world, to the measurement results, which in this case would be determined as the result of the experiment $\mathbb{M}$.

The function $d_{\mathbb{M}}$ establishes a connection between what a theory of consciousness T predicts or postulates about physical events in the real world, on the one hand, and the possible measurement outcomes that can occur according to T, on the other. It selects from the events $\mathcal{P}_T$ that the theory T is committed to those events which determine the data that is stored on D. Making use of the symbol $\mathcal{O}_T$ introduced above to denote the possible measurement results that can occur in $\mathbb{M}$ after T, this means that

$$d_{\mathbb{M}}(\mathcal{P}_T) = \mathcal{O}_T \,. \tag{7}$$

In this way, we can determine $\mathcal{O}_T$ independently of approximations or auxiliary assumptions.

6. Why the closure of the physical is unscientific

By considering that measurement results must be stored and are thereby determined by physical events, we have obtained a novel, additional handle for analyzing experiments in the scientific study of consciousness. In addition to what experimenters derive from a theory T and appropriate auxiliary assumptions, we can now analyze measurement results along the path of what a theory of consciousness says about physical events. This gives rise to the following theorem.

Theorem 1. *The closure of the physical is unscientific.*

Proof. Let T_1 and T_2 denote two theories of consciousness which obey the closure of the physical. This implies that there exist embeddings $\iota_1 : \mathcal{P}_{T_1} \longrightarrow \mathcal{P}_P$ and $\iota_2 : \mathcal{P}_{T_2} \longrightarrow \mathcal{P}_P$ as in (1). Let $\mathbb{M}$ denote an experiment to which both T_1 and T_2 are applicable, and D the data storage device(s) used in that experiment. Because of condition (4), we have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\mathcal{P}_{T_1})$ and $d_D(\iota_2(\mathcal{P}_{T_2})) = d_D(\mathcal{P}_{T_2})$.

Both T_1 and T_2 need to be committed to the existence of physical events which instantiate the data storage device D used in $\mathbb{M}$, for otherwise they would violate the very conditions that make $\mathbb{M}$ possible. Therefore, D is instantiated in both $\mathcal{P}_{T_1}$ and $\mathcal{P}_{T_2}$. Because applying ι_1 resp. ι_2 does not change the content of the described events, it follows that D is also instantiated $\iota_1(\mathcal{P}_{T_1})$, resp. $\iota_2(\mathcal{P}_{T_2})$.

Because ι_1 is an embedding, we have $\iota_1(\mathcal{P}_{T_1}) \subset \mathcal{P}_P$. Because D is instantiated in $\iota_1(\mathcal{P}_{T_1})$, Equation (5) applies so that we have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\mathcal{P}_P)$. The same applies to ι_2, so that also here, Equation (5) implies $d_D(\iota_2(\mathcal{P}_{T_2})) = d_D(\mathcal{P}_P)$. So we in fact have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\iota_2(\mathcal{P}_{T_2}))$, which in light of the above implies $d_D(\mathcal{P}_{T_1}) = d_D(\mathcal{P}_{T_2})$.

We thus find that the data stored on D is exactly the same for both theories. Restriction to $d_{\mathbb{M}}$ introduced in (6) furthermore implies that $d_{\mathbb{M}}(\mathcal{P}_{T_1}) = d_{\mathbb{M}}(\mathcal{P}_{T_2})$, and because of (7), this implies that $\mathcal{O}_{T_1} = \mathcal{O}_{T_2}$. So the measurement results of $\mathbb{M}$ are

exactly the same according to both T_1 and T_2. Independently of which predictions one arrives at by making use of auxiliary assumptions, the closure of the physical implies that the data sets which can occur in $\mathbb{M}$ cannot differ.

Since $\mathbb{M}$ was chosen arbitrarily, this conclusion holds for any experiment $\mathbb{M}$, so T_1 and T_2 are empirically indistinguishable. And because T_1 and T_2 were arbitrarily chosen among the theories obeying the closure of the physical, we can conclude that all theories obeying the closure of the physical are empirically indistinguishable. It follows that the closure of the physical is an assumption that is unscientific. □

7. Causal closure of the physical

The *causal closure of the physical* is the assumption that for every physical effect there is a sufficient physical cause. This is an ontological assumption; it refers to what is the case in the actual world. In contrast, the assumption we have been working with above – that a theory of consciousness obeys the *closure of the physical* if and only if it does not postulate changes in physical events explained, predicted, or otherwise determined by natural science – is epistemic in nature, it depends on the definition, formulation, and content of a theory of consciousness.

The precise meaning of the causal closure of the physical depends heavily on what notion of causality one subsumes, what ontology one grants to causality (if any), and what one allows as relata of the causal relation. Nevertheless, there is a great deal of consensus about what epistemic implications this assumption has.

According to Jaegwon Kim, for example, the causal closure of the physical implies that "to explain the occurrence of a physical event we never need to go outside of the physical realm" [12, p. 147]. And Frank Jackson characterizes the causal closure of the physical as the claim that "the physical sciences, or rather some natural extension of them, can in principle give a complete explanation for each and every bodily movement, or at least can do so up to whatever completeness is compatible with indeterminism in physics" [11, p. 378].

These statements exemplify that the causal closure of the physical is generally taken to imply that every physical event which is explained at all, is explainable by natural science. But explanation, precisely construed [22], is only one way in which a theory can address events. Making room for prediction and other possible ways as well, we may take the causal closure of the physical to imply that every physical event which is predicted, explained, or determined at all, can be predicted, explained, or determined by natural science.

Applied to a theory of consciousness, this means that any physical event that the theory explains, predicts, or determines can eventually be explained, predicted, or determined by natural science. But for this to be true, the theory must not replace, alter, or add to the natural science account of physical events, because otherwise it

would be committing itself to physical events that cannot be explained, predicted, or determined by natural science. Thus, the causal closure of the physical implies that a theory of consciousness cannot make changes to the physical events that are explained, predicted, or determined by natural science.

This point can also be stated in formal terms. We have denoted the set of physical events that a theory of consciousness is committed to by $\mathcal{P}_T$. These are the events explained, predicted, or otherwise determined by that theory. And we have denoted the set of physical events explained, predicted, or otherwise determined by natural science (now or in the future) by $\mathcal{P}_P$. Thus, if every physical event that can be explained, predicted, or determined at all can be explained, predicted, or determined by natural science, then every event that is in $\mathcal{P}_T$ is also in $\mathcal{P}_P$. Taking into account the different languages that can be used in the two cases, this means that for every event description in $\mathcal{P}_T$ there is a corresponding event description of the same event in $\mathcal{P}_P$. This constitutes an injective function that maps $\mathcal{P}_T$ to $\mathcal{P}_P$.

We thus arrive at exactly the same formal requirement as in Equation (1). The causal closure of the physical implies that there is an embedding $\iota : \mathcal{P}_T \to \mathcal{P}_P$ that specifies for each physical event and physical description that the theory of consciousness is bound to the corresponding event and description explained, predicted, or determined by natural science. Causal closure of the physical implies closure of the physical, and as a corollary of Theorem 1 we posit that causal closure of the physical is also unscientific.

We emphasize that nowhere in our argument do we restrict to physical events which are already explained or predicted by natural science. What matters is only which relation a theory of consciousness proposes between the physical events it is committed to and the physical events that natural science posits. Even if a theory presupposes that the physical events it associates with conscious experiences are determined by physical laws, but cannot in practice be explained or predicted based on these laws, as some emergentist theories would have it, our argument applies. Theories of this sort may be wrong about what they say about physical events, and experiments may help to determine whether this is the case, but insofar as they buy into the very same underlying account of physical events as all other theories, the measurement results necessarily are the same as if any other theory were true.

8. Conclusion

We have shown that the causal closure of the physical goes far beyond what is usually considered. Since all measurement results in the scientific study of consciousness are either physical events (such as keystrokes or sound waves) or at least determined by physical events (such as data stored on hard disks), no two theories obeying the causal closure of the physical can actually be distinguished in experiments. Our result applies to all major neuroscientific theories of consciousness as well as to the leading

philosophical paradigms in the field. It applies to any theory of consciousness that fits into the natural science account of physical events without altering it. This includes all functionalist and identity theories of consciousness, such as GNW [14], HOT [1], AST [8], or predictive processing-based theories [20], as well as eliminativist or illusionist theories [6]. But it also includes theories such as IIT, whose mathematics takes the form of a function that maps physical states and events to conscious states and events [13].

We have shown that no experiment of any kind can actually distinguish between these theories. Whatever measurement result is consistent with one theory is necessarily consistent with the other, because the physical functioning of the brain, from stimulus presentation to verbal message or similar output, is exactly the same according to all these theories. This observation is at odds with the numerous experiments conducted to date to distinguish precisely between some of these theories. Our results show that there is a major flaw underlying *all* of these experiments. The theories on which these experiments are based violate a necessary condition for the experiments to work as intended.

At best, we can conclude that experimenters do not really adhere to the closure of the physical when conducting experiments, so they implicitly assume that the theories tested modify what falls solely within the realm of natural science. At worst, our results call into question the conclusions drawn on the basis of these experimental results. In either case, our results show that the closure of the physical must be abandoned in both theory and experiment. Theories of consciousness must explicitly state how what they take to be consciousness (physical or otherwise) comes to determine reports and other measures of consciousness, and to do this they must enter the realm of natural science.

In a very different context, Einstein once asserted that "[it] is the theory which decides what we can observe" [5, 9]. It seems that this point has not yet been fully recognized in the construction of scientific theories of consciousness.

Acknowledgments. We gratefully acknowledge support from the Foundational Questions Institute (FQXi) and the Fetzer Franklin Fund. Thanks also to Sander Beckers, Alexander Gebharter, Kobi Kremnitzer, Christian List, and Wanja Wiese for helpful discussions, and to Joe Dewhurst, Timo Freiesleben, and Naftali Weinberger for feedback on an earlier draft.

References

[1] R. Brown, H. Lau, and J. E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019.

[2] D. J. Chalmers. How can we construct a science of consciousness? In *The Cognitive Neurosciences III*, pages 1111–1119. MIT Press, Boston, MA, 2004.

[3] D. J. Chalmers and K. J. McQueen. Consciousness and the collapse of the wave function. *arXiv preprint arXiv:2105.02314*, 2021.

[4] A. Del Cul, S. Baillet, and S. Dehaene. Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, 5(10):e260, 2007.

[5] T. Filk. It is the theory which decides what we can observe (einstein). In *Contextuality from Quantum Physics to Psychology*, pages 77–92. World Scientific, Singapore, 2016.

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

[7] S. Frässle, J. Sommer, A. Jansen, M. Naber, and W. Einhäuser. Binocular rivalry: frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience*, 34(5):1738–1747, 2014.

[8] M. S. Graziano and T. W. Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500, 2015.

[9] W. Heisenberg. *Physics and Beyond*. Allen & Unwin, London, 1971.

[10] E. Irvine. Measures of consciousness. *Philosophy Compass*, 8(3):285–297, 2013.

[11] F. Jackson. Mental causation. *Mind*, 105(419):377–413, 1996.

[12] J. Kim. *Philosophy of Mind*. Westview Press, Boulder, 1996.

[13] J. Kleiner and S. Tull. The mathematical structure of integrated information theory. *Frontiers in Applied Mathematics and Statistics*, 6:74, 2021.

[14] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

[15] T. Metzinger. The problem of consciousness. In T. Metzinger, editor, *Conscious Experience*, pages 3–37. Imprint Academic/Schoningh Thorverton, UK, 1995.

[16] M. Oizumi, L. Albantakis, and G. Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, 2014.

[17] D. Papineau. The causal closure of the physical and naturalism. In A. Beckermann, B. P. McLaughlin, and S. Walter, editors, *The Oxford Handbook of Philosophy of Mind*, pages 53–65. Oxford Universaity Press, Oxford, 2009.

[18] M. Pauen. Painless pain: Property dualism and the causal role of phenomenal consciousness. *American Philosophical Quarterly*, 37(1):51–63, 2000.

[19] M. Pauen. Feeling causes. *Journal of Consciousness Studies*, 13(1-2):129–152, 2006.

[20] T. Schlicht and K. Dolega. You can't always get what you want. *Philosophy and the Mind Sciences*, 2, 2021.

[21] M. Sprevak and E. Irvine. Eliminativism about consciousness. In A. Beckermann, B. P. McLaughlin, and S. Walter, editors, *Oxford Handbook of the Philosophy of Consciousness*, pages 348–370. Oxford University Press, Oxford, 2020.

[22] M. Strevens. Scientific explanation. In D. Borchert, editor, *Encyclopedia of Philosophy*, pages 518–27. Macmillan Reference USA, New York, 2006.

[23] N. Tsuchiya, M. Wilke, S. Frässle, and V. A. Lamme. No-report paradigms: extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*, 19(12):757–770, 2015.

[24] J. O. Weatherall. Theoretical equivalence in physics, Part 1. *Philosophy Compass*, 14(5):e12592, 2019.

[25] J. O. Weatherall. Theoretical equivalence in physics, Part 2. *Philosophy Compass*, 14(5):e12591, 2019.