MODELS OF CONSCIOUSNESS

JOHANNES KLEINER†

†Institute for Theoretical Physics, University of Hannover

“Mathematics translates concepts into formalisms and applies those formalisms to derive insights that are usually not amenable to a less formal analysis.” [Jos15]

Abstract. The scientific study of consciousness is a new field which has emerged as a response to groundbreaking developments in neuroscience, cognitive psychology and analytic philosophy. Its aim is to develop a scientific account, formulated in terms of formal mathematical laws or regularities, of how conscious experience relates to the physical domain (e.g. to brain processes). Even though several models of consciousness have been proposed in the literature, details of the underlying conceptual terminology and formal framework remain poorly understood. This paper aims to improve this situation by proposing a grounding of the scientific study of consciousness which is based on a concise definition of the object of investigation and gives rise to a clear-cut mathematical framework for models of consciousness. Several examples are studied and an application of this framework is given. The underlying definitions are a further development of the grounding of the scientific study of consciousness developed by David Chalmers and Thomas Nagel.

Contents

1. Introduction 2
2. Chalmers’ Grounding of the Scientific Study of Consciousness 5
3. Phenomenological Grounding of the Scientific Study of Consciousness 7
  3.1. Basic Definitions 7
  3.2. Explanatory Gap 21
  3.3. Theories, Models, Symmetries 22
  3.4. Models of Consciousness 23
  3.5. Comparison with Direct Description 28
  3.6. Closure of Physical 30
4. Examples 31
  4.1. Integrated Information Theory 31
  4.2. Global Neuronal Workspace Theory 34
  4.3. Conscious Agent Networks 37
  4.4. Expected Float Entropy Minimisation 39
5. Application 41
  5.1. Novel Symmetries 42
  5.2. Qualia Determined Locally 46
6. Conclusion & Outlook 55
Appendix A. Conceptual Problems of Chalmers’ Grounding 56
References 63
1. Introduction

Any research activity associated with the scientific study of consciousness implicitly or explicitly refers to what may be called a ‘grounding’ of the scientific study of consciousness:

**Definition 1.1.** A grounding of the scientific study of consciousness contains at least
- an explicit definition of what is to be studied.
- an explicit outline of the methodology.

Here “methodology” refers to “a collection of methods, practices, procedures and rules used by those who work in some field” [Wik18b], “a system of methods used in a particular area of study or activity” [Oxf18]. In particular, the methodology includes the specification of what constitutes an experiment.

Over the last two decades, the scientific study of consciousness has largely been guided by a grounding developed by David Chalmers in [Cha96]. This grounding has played a pivotal role in the creation and consolidation of the field. However, it also exhibits several severe problems (Appendix A). This paper introduces an alternative to Chalmers’ grounding. The main motivation, which has led to this proposal, was to clarify which formal structures are to be used when building models of consciousness, i.e. models of the mind-matter relation. As shown by this work, this question can be answered largely without specific metaphysical commitments. A nice – though unintentional – feature of this grounding is that it also avoids all major conceptual problems of Chalmers’ grounding.

The starting point of this alternative grounding is to turn a key characteristic of what is generally referred to as qualia or phenomenal consciousness into a definition: That these aspects of experience cannot be collated, i.e., that the identity of qualia of two different experiencing subjects cannot be determined. We refer to corresponding aspects of experience simply as non-communicable aspects of experience:

**Definition 3.8.** We define the term *qualia* to refer to all aspects of experience of an experiencing subject which cannot be communicated to other experiencing subjects in some class $C$.

This is a phenomenological definition, which refers to experience as “we” find it. Whether the non-communicability of some aspect of experience is in some sense fundamental or a result of particularities of a system’s architecture is a question to be answered a posteriori by models of consciousness. To some extent, this definition reflects that there are “seemingly ineffable qualitative properties of phenomenology” [Ton08, p. 229], as well as the purported fact that “qualia are ineffable – in fact the paradigm case of ineffable items” [Den93, p. 384], that “[w]e have no independent language for describing phenomenal qualities (...), there is something ineffable about them” [Cha96, p. 22], and that “there are facts that do not consist in the truth of propositions expressible in a human language” [Nag74, p. 441].

The aspects of experience which satisfy Definition 3.8 are of special interest because the non-communicability induces a fundamental difficulty in any scientific approach: The non-communicability implies that these aspects cannot be referenced intersubjectively, which in turn implies that they cannot be referenced in a scientific model or

---

1 The term ‘grounding’ is one of several translations of the German word “Grundlegung”.
2 “Abgleichen” in German.
empirical analysis. There is a fundamental explanatory gap (Section 3.2). The goal of this paper is to develop a framework, a “new theoretical form” [Nag74, p. 436], which takes this non-communicability into account and allows to define meaningful scientific models which address both communicable and non-communicable aspects of experience.

To this end, we first specify how basic phenomenological facts can be used to construct a mathematical space $L$ related to experience. This is a space of labels, whose structure represents communicable relations between non-communicable aspects of experiences. The automorphism group of this space (the group of structure-preserving bijections) precisely quantifies the ambiguity of any statement in terms of qualia: The intersubjectively meaningful references to qualia are the equivalence classes induced by this group. These equivalence classes are susceptible to the usual scientific explanation. The task of a scientific study of consciousness is to go beyond these: To propose and empirically test models of consciousness which also have something to say about qualia sensu stricto, i.e. individual non-communicable aspects of experience.

Based on this space $L$ we define pre-models of consciousness as a unification of a physical theory $T_P$ with multiple copies of this space (one for each experiencing subject that is being modelled). Here, the prefix ‘pre’ indicates that pre-models of consciousness are not yet well-defined: Any reference to labels, e.g. in a trajectory of such a unified theory, still carries the ambiguity induced by the non-communicability. To remedy this situation, we show that a necessary and sufficient condition for pre-models of consciousness to be empirically well-defined is that they carry a specific symmetry. The corresponding symmetry group is the automorphism group of the space $L$, denoted by Aut($L$). Therefore, models of consciousness need to be defined as:

**Definition 3.28.** A model of consciousness is a pre-model of consciousness $M$ (Definition 3.24) which carries an Aut($L$)$^k$ symmetry of the form (3.22).

Crucially, the form (3.22) allows for an action of this group on the physical domain. In Section 3.5, we show that this seemingly minor detail is what makes the difference to a direct description of qualia and physical states, and what allows a model of consciousness so defined to address qualia sensu stricto in a meaningful way.

The grounding so developed is metaphysically neutral: Given a description of experience and a physical theory of choice, models of consciousness can be constructed which express varying metaphysical positions, ranging from functionalisms and physicalisms to idealisms, panpsychisms, dualisms and dual aspect monisms. In particular, unlike Chalmers’ grounding (Section 2), this grounding does not need to assume the closure of the physical in order to identify those aspects of conscious experience which are being studied. Correspondingly, models of consciousness can be constructed which describe the physical as closed (Section 3.6) just as well as models which do not describe the physical as closed. (Note, however, that a transcendental argument, which is given in Remark A.1, shows that assuming the closure of the physical violates a necessary condition of the possibility of a scientific study of consciousness.) Examples from the literature, based on different metaphysical positions, are studied in Section 4.

Since the mathematical structure of the space $L$ is grounded in experience, and hence ultimately needs to be justified by a phenomenological analysis, this alternative
grounding of the scientific study of consciousness is dubbed *phenomenological grounding*. It is a further development of David Chalmers’ previous work \cite{Chal96, Chal10} and largely inspired by the characterisation of experience given by Thomas Nagel \cite{Nag74}. In particular, the introduction of models of consciousness as unified descriptions of a physical theory and spaces \( L \) is a generalization, within this grounding, of what can be called ‘Chalmers’ strategy’ to address an explanatory gap. Problems which arise if one wishes to apply Chalmers’ grounding (Appendix A) have also played an important role in the choice of definitions.

In Section 5 we return to the symmetry that a model of consciousness needs to carry. In many scientific theories, the existence of symmetries has far-reaching consequences. The task of Section 5 is to study whether this is the case here as well.

We have noted above that a model of consciousness constitutes a unified description of a physical theory with the space \( L \). Let \( M \) denote a model of consciousness and let \( T_P \) denote the physical theory it is based on. In Section 5.1 we ask under which circumstances the \( \text{Aut}(L) \) symmetry of \( M \) is also a symmetry of \( T_P \) and vice versa. To answer this question, we prove three lemmas: First, Lemma 5.2 states that if \( M \) describes the physical as closed, it follows that \( \text{Aut}(L) \) restricts to a symmetry of \( T_P \). Since the physical theory \( T_P \) is assumed to be a well-known (well-studied) physical theory, this lemma says that any model of consciousness, which describes the physical as closed, carries a known symmetry of the physical domain. Lemma 5.3 and Lemma 5.4 show the opposite case. They state that if \( M \) does not describe the physical as closed and satisfies a technical assumption, it follows that \( \text{Aut}(L) \) does not restrict to a symmetry of \( T_P \). These lemmas thus prove that any model of consciousness which does not describe the physical as closed carries a novel, previously unknown symmetry compared to \( T_P \). This surprising relationship, which is rooted in the mathematical definition of symmetries in modern physics, is of particular relevance in light of Remark A.1 mentioned above. The technical assumptions of both lemmas are summarized at the beginning of Section 5.1.

In Section 5.2, we further investigate the latter case. To this end, we consider models of consciousness which are based on fundamental physical theories such as quantum theories of particles or fields, or classical theories such as electromagnetism or gravitation. The construction of any such model is most likely a difficult task, at least if one wishes to take into account basic physical principles such as e.g. the principle of relativity. However, due to the \( \text{Aut}(L) \) symmetry, something can be said about any reasonable model of this kind.

To explain this, we recall again that a model of consciousness \( M \) is defined to be a unified description of a physical theory \( T_P \) and \( L \). Therefore, the laws of \( M \) are generalizations of the laws of \( T_P \): They describe both the physical domain and consciousness. Generally speaking, this means that the laws of \( T_P \) should in some mathematical sense be contained in the laws of \( M \). Together with the fact that the laws of the above-mentioned fundamental physical theories all contain derivative operators with respect to spatial coordinates, this motivates us to consider models of consciousness whose laws contain such a derivative operator as well. If this is the case, Proposition 5.10

\[ \text{If neurons are but complex configurations of fundamental physical quantities (molecules, atoms or quantum fields), any law or regularity which describes how experience relates to neural activity should eventually be cast in a language compatible with the fundamental physical quantities.} \]
and Corollary 5.12 apply, the latter of which states the following. Note that all terms used in this corollary are defined in Sections 5.2.1 to 5.2.4.

**Corollary 5.12.** Any non-trivial model of consciousness $M$ which

1.) contains invariant spatial derivatives,
2.) does not postulate the physical closed,
3.) satisfies the technical assumptions of Lemma 5.3 or Lemma 5.4,
4.) and determines qualia locally

necessarily contains a field (5.14) which is not part of the physical theory $T_P$ but interacts with the dynamical quantities of the latter.

In simple terms, this corollary thus states that if the Assumptions 1 to 4 hold, it follows that $M$ necessarily contains a new physical field compared to $T_P$. This is very similar to – and in fact motivated by – the well-known ‘gauge principle’ of physics, which establishes the existence of new interacting fields in response to new symmetry requirements. The symmetry in question here is the Aut($L$) symmetry of $M$. Since the field necessarily interacts with the dynamical quantities of the physical theory $T_P$, observational consequences will generally follow once Corollary 5.12 is applied in a process of constructing models of consciousness.

Section 5 thus offers a first application of the phenomenological grounding. Since this application depends neither on particularities of the mathematical structure of the space $L$, nor on the specific form of the proposed laws of a model of consciousness, it shows that new physical predictions are not only possible but in fact to be expected. The next step of the programme suggested by this result is the construction of specific models of consciousness which are based on fundamental physical theories, guided by metaphysical positions on the mind-body problem and constrained by neuroscientific findings, so as to eventually be able to make novel experimental predictions.

> "Many scientific discoveries have been delayed over the centuries for the lack of a mathematical language that can amplify ideas and let scientists communicate results." [Pea09, p. 427]

### 2. Chalmers’ Grounding of the Scientific Study of Consciousness

The most prominent grounding of the scientific study of consciousness has been developed by David Chalmers in [Cha96]. Since it is the basis of the grounding proposed in what follows, we review its essential definitions. Note, however, that the following outline of Chalmers’ grounding is intended to highlight the relations among various constituents of his grounding and is not intended to be of an introductory nature. A good and short introduction to this topic is [Cha10, Ch. 1].

First, we note that Chalmers’ definition of ‘physical domain’ includes what is often called ‘material’ or ‘physical’ configurations, such as neurons or brain tissue, as well as more fundamental physical notions such as “mass, charge, and space-time” [Cha10, p. 17] or “atoms, electro-magnetic fields, and so on” [Cha96, p. 71]. We thus define the term ‘physical domain’ to refer to all those phenomena which are currently considered to be the subject of a natural science (physics, chemistry, earth science, biology, etc. [Wik18a]). Chalmers assumes that:

(A1) “The physical domain is causally closed.” [Cha96, p. 161]

“For every physical event, there is a physical sufficient cause.” [Cha96, p. 125]
Central to Chalmers’ grounding are the terms ‘function’ and ‘structure’. “Here ‘function’ is not used in the narrow teleological sense of something that a system is designed to do but in the broader sense of any causal role in the production of behaviour that a system might perform” [Cha10, p. 6]. The term ‘structure’ is used in a spatiotemporal sense. Together, they constitute, according to Chalmers, the notion of explanation which is used throughout contemporary science: “One can argue that by the character of physical explanation, physical accounts explain only structure and function, where the relevant structures are spatiotemporal structures, and the relevant functions are causal roles in the production of a system’s behavior.” [Cha10, p. 105f.]

We denote this notion of explanation by $\text{(E1)}$. Assuming some laws or theories relating to the physical domain as given (= accepted by the scientific community by and large) and referring to them as ‘accepted theoretical notions’, $\text{(E1)}$ might be put as follows:

$\text{(E1)}$ An explanation specifies the function and structure of an explanandum in terms of the function and structure of accepted theoretical notions.

The crucial aspect of Chalmers’ grounding is to establish, in a consistent and explicit way, that there are phenomena, related to consciousness, to which no function or structure (as defined above) can be associated. It follows that these phenomena cannot be explained according to $\text{(E1)}$ and hence, if $\text{(E1)}$ indeed captures all notions of explanations which are used throughout contemporary science, that they cannot be explained by contemporary science. – There is an “explanatory gap” [Lev83, Cha96]. Chalmers refers to these phenomena as “phenomenal concepts”, “phenomenal qualities” or “qualia” [Cha96]. We refer to these phenomena as ‘phenomenal aspects of consciousness’:

$\text{(D1)}$ Phenomenal aspects of consciousness are those aspects of conscious experience which do not have a function or structure, where ‘function’ and ‘structure’ are as defined above.

The key requirement for this definition of what is to be studied by a science of consciousness to make sense is to establish that there are aspects of experience which satisfy $\text{(D1)}$, i.e. which neither have a spatio-temporal structure nor a causal role in the production of behaviour. It is the second requirement with respect to which $\text{(A1)}$ is crucial, for $\text{(A1)}$ can be utilized to argue that nothing non-physical can have a causal influence on the physical domain. Therefore, all aspects of experience which do not have a spatio-temporal structure (e.g. in the Cartesian sense of being non-extended in space and space-time) automatically satisfy $\text{(D1)}$. We will not review the various arguments which aim to prove the existence of phenomenal aspects of consciousness at this point.

Put in terms of Definition 1.1, what is to be studied in the scientific study of consciousness are, according to this grounding, phenomenal aspects of consciousness and their relation to the physical domain. Since these are, by definition, not accessible to the usual scientific methodology, Chalmers proposes that the task of a science of consciousness is to find what he calls “psychophysical laws” [Cha96, p. 127] which

$^4$The term “hard problem of consciousness” [Cha95] is also used in this context.

$^5$In [Cha10], he prefers to use the term ‘experience’: “Sometimes terms such as ‘phenomenal consciousness’ and ‘qualia’ are also used here, but I find it more natural to speak of ‘conscious experience’ or simply ‘experience.’” [Cha10, p. 5].
relate the physical domain to phenomenal aspects of consciousness. Due to Assumption (A1) and an underlying stance on the nature of causality “[t]hese laws will not interfere with physical laws; physical laws already form a closed system. Instead, they will be supervenience laws, telling us how experience [= phenomenal aspects of consciousness] arises from physical processes” [Cha96, p. 127]. In combination with (E1), this implicitly points at the major parts of the methodology to be used according to this grounding.

Chalmers’ grounding raises several questions related to the definition and ontological status of causality, to the validity of Assumption (A1), to the nature of experiments in his grounding and to the validity of the subsumed notion of explanation, which we discuss in Appendix A. The upshot is that there are severe conceptual problems which make it questionable whether a scientific research programme based on this grounding can be carried out at all.

Furthermore, any scientific approach based on this grounding faces the question of which mathematical structure one is to use in order to describe phenomenal aspects of consciousness when formulating “psychophysical laws” [Cha96, p. 127]. Whereas the physical domain comes with a clear-cut mathematical structure, Chalmers’ grounding merely asserts that the phenomenal aspects form a set and offers no systematic way of tying additional mathematical structure to the phenomenology of experience.

This strongly suggest the construction of other groundings of the scientific study of consciousness. In the remainder of this article, we introduce and apply such an alternative which avoids the above-mentioned problems. Whereas this grounding breaks with several of Chalmers’ main ideas, it retains the key idea of addressing an explanatory gap with mathematical tools. Similar to Chalmers’ grounding, this grounding is largely inspired by a particular reading of the ground-breaking article [Nag74].

3. Phenomenological Grounding of the Scientific Study of Consciousness

3.1. Basic Definitions. The starting point of every scientific activity related to consciousness is a preliminary choice of a class \( C \) of experiencing subjects. The object of investigation of any empirical study, and what informs any model-building process, is experience in the following sense.

Definition 3.1. We use the term ‘experience’ to denote the totality of impressions, feelings, thoughts, perceptions, etc. which an experiencing subject lives through at a particular instant of time.

The general idea underlying any conception of the scientific study of consciousness is to study experience and its relation to the physical domain by scientific means. Mostly, some part or feature of experience is under consideration. In order to emphasise that this part or feature may not be strictly separable from other parts of features, we use the term ‘aspects of experience’:

Definition 3.2. Aspects of experience denote specific or general features, parts or elements of a particular experience or of a set of experiences.

6Two remarks are in order. First, no special focus on subjectivity is intended when using the term ‘experiencing subject’. Alternatively, one could use the term ‘experiencer’. Second, we remark that the meaning of ‘instant’ is to be fixed during the model-building process. It could refer to physical just as well as to experiential instants of time.
Example 3.3. Aspects of experience range from individual visual, auditory or tactile experiences to general characteristics, such as the experience of a first person perspective, the unity of the conscious scene \cite{Set07}, or the structure and composition of experience \cite{OAT14}.

A priori, every experiencing subject only has access to his/her own experience. However, systematic investigations of which aspects of experience are invariant over a large class of experiencing subjects are possible and have been carried out since the early 20th century as part of the philosophical discipline of phenomenology.

**Definition 3.4.** A *phenomenological fact* is a statement about aspects of experience which holds for all experiencing subjects in a class $\mathcal{C}$.

Phenomenological facts serve as a starting point for any investigation in the scientific study of consciousness. In empirical studies, they are what can be correlated with physical states, e.g. to construct neural correlates of consciousness. When building models of consciousness, they are what informs the choice of mathematical structure.

Next, we make use of three basic phenomenological facts in order to examine in more detail which methodology may be used to study aspects of experience and their relation to the physical domain.

**Phenomenological Fact 3.5.** Aspects of experience can be divided into two classes:

a) Aspects of experience which cannot be collated, i.e. whose identity over several different experiencing subjects in the class $\mathcal{C}$ cannot be determined. These are, in short, experienced as *non-communicable*.

b) Aspects which can be collated, i.e. which are experienced as *communicable* to other experiencing subjects in the class $\mathcal{C}$.

The former include aspects which are frequently referred to as having a “subjective character” \cite[p. 437]{Nag74}, connected to a “particular point of view” \cite[p. 441]{Nag74}, or “experienced as private”. The latter include aspects which are experienced as accessible also “from other points of view” \cite[p. 443]{Nag74} or as having an “objective nature” \cite[p. 443]{Nag74}.

**Remark 3.6.** When being presented with Phenomenological Fact 3.5, scientists usually tend to think about how this can be derived from a theory of language. In our opinion, the more important task is to ground the underlying distinction in a proper phenomenological analysis. This is so because, in simple terms, phenomenological facts are statements about how experiencing subjects, such as the reader or the author, find themselves experiencing. One could say that they express invariant facts of ‘what experience is like’ or of how experience ‘reveals itself’.

Since this phenomenological fact may, at first, be counter-intuitive, we illustrate it in a simple example.

\footnote{The restriction to a class $\mathcal{C}$ of experiencing subjects is necessary because a phenomenological analysis of invariants of experience is always restricted to experiencing subjects which are similar in some respects: “[O]ne person can know or say of another what the quality of the other’s experience is. [However, this] ascription of experience is possible only for someone sufficiently similar to the object of ascription to be able to adopt his point of view” \cite[p. 442]{Nag74}. However, the choice of class $\mathcal{C}$ is not a constraint for models of consciousness, but rather a starting point, i.e. a preliminary choice which informs the model-building process. Models may eventually allow to determine which organisms experience. We note also that the name ‘phenomenological fact’ is a reference to phenomenology rather than an attempt to condense the phenomenological method into a simple definition.}
Example 3.7. Most aspects of experience are communicable. To give a simple example, consider an experiencing subject in an artist’s workshop and assume the experiencing subject is looking at the artist’s canvas, perceiving its rectangular shape, the scaffold, some brushes, the background as well as some unfinished painting. All of the items just mentioned are part of the subject’s experience and are experienced as communicable: We can meaningfully communicate to other experiencing subjects about our experience containing a canvas, its shape, a scaffold, brushes, etc. The communicability of these aspects of the experience can be taken as a plain fact about how experiencing subjects (in the class $C$ chosen above) find themselves in the world.

A simple and useful example of aspects of experience which are not communicable are colour aspects of experience. Whereas the experiencing subject could communicate differences in colours on the artist’s canvas, there simply are no means to communicate the particular colour the subject is experiencing when a particular mixture of wavelengths impinges on his/her eyes. This simple but important fact is pointed at by the plain question, sometimes already raised during elementary school, of how two experiencing subjects might come to conclude that the experience of colour which they have if they look at, e.g., the clear sky is the same. They may ensure that they use the same reference (‘blue’) for the experience, that they see the same wavelength and they might even be able to conclude that similar neuronal assemblies are active in both of their brains while having the experience in question. However, none of this is a priori related to the colour aspects of their experiences (‘what it is like to see blue’).

There is at present no possibility to even meaningfully ask the question of whether colour experiences of two experiencing subjects are the same or different. There is no reasonable way to assign truth values to statements of the form ‘my colour experience $f_1$ is equal to your colour experience $f_2$’, equality is not a well-defined concept when referencing to experiences of two different experiencing subjects. I.e., colour experiences are non-communicable aspects of experience in the sense of Phenomenological Fact 3.5.

This non-communicability has consequences for any scientific account of colour experience. E.g., any hypothesis that a particular neural activity occurs whenever a subject is experiencing a colour ‘green’ is not well-defined, simply because there is no intersubjectively meaningful reference to ‘green’: the colour experience one subject is having when when presented a 510nm light source may be very different from the colour experience another subject is having when presented the same light source. In other words, any intersubjective reference to colour experiences carries a certain ambiguity, which has to be taken into account when constructing models or designing experiments related to colour experience.

The main point of this paper, argued for in detail below, is that non-communicable aspects of experience cannot be addressed by the usual scientific methodology. Since the term ‘qualia’ is generally used to denote what is considered as essential in a particular analysis of experience, we introduce the following abbreviation.

Definition 3.8. We define the term *qualia* to refer to all aspects of experience of an experiencing subject which cannot be communicated to other experiencing subjects in the class $C$.

Example 3.9. According to Example 3.7, colour experiences satisfy the condition of Definition 3.8. Thus colour experiences are qualia.

Example 3.10. Example 3.9 is a special case of the aspects of experience referenced by Thomas Nagel in [Nag74] when introducing his famous notion of ‘What is it like to be ...?’:

---

8Note that the description of this example is itself an instance of such communication.
9This ‘plain fact’ may eventually change as a result of changes in society (including science).
10Eventually, one might wish to choose the class $C$ as small as possible for a particular scientific task, cf. Remark 3.32.
11We generally abbreviate ‘colour aspects of experience’ by ‘colour experience’.
“[F]undamentally an organism has conscious mental states if and only if there is something that it is like to be that organism – something it is like for that organism. We may call this the subjective character of experience.” (p. 436)

Nagel also uses the term “how it is for the subject himself” (p. 440) to point to these aspects of experience. Though Nag74 does not make the distinction of Phenomenological Fact 3.5 central to his line of reasoning, one can find hints toward this distinction in Nag74. E.g., he claims that “we do not possess the vocabulary to describe [what it is like to be us] adequately” (p. 440), there are “facts that do not consist in the truth of propositions expressible in a human language.” (p. 441)

The goal of the scientific study of consciousness, according to this grounding, is to investigate how both communicable and non-communicable aspects of experience relate to the physical domain. With respect to the former, this can be achieved by the usual scientific methodology. Since communicable aspects of experience can be referenced in an empirical analysis or model building process just as any other explanandum in science, there is no fundamental obstacle involved. Qualia, on the other hand, are by definition aspects of experience which cannot be collated. This implies that they also cannot be referenced (individually) in an empirical study or in a formal model of consciousness. This poses a fundamental problem for the usual scientific methodology. How can something be studied empirically which is only accessible from one point of view?

In what follows, we will show that this problem is less severe than one might assume. First, in Section 3.1.2, we show that non-communicable aspects can be represented by a mathematical form which takes into account the non-communicability and allows to specify precisely the ambiguity involved in any reference to qualia. In Section 3.4, we show that this mathematical representation of experience allows to construct models of consciousness of various kinds which allow to address, in an empirically well-defined way, individual qualia. The hope of this grounding, and of the scientific study of consciousness by and large, is that some of these models can eventually be subjected to the same form of theory confirmation or falsification which is in use in other areas of science.

3.1.1. Formal Representation of Experience. In order to define a formal representation of experience, we make use of two further basic phenomenological facts. These are very general in nature and it is plausible that they hold independently of the particular choice of class C. However, due to the restricted possibility of phenomenological analysis mentioned above, we generally assume C to comprise adult humans. The first phenomenological fact expresses the observation that some qualia are experienced as identical, whereas others are not, or in other words, that one sometimes experiences a non-communicable aspect as identical to a non-communicable aspect one has experienced at another time.

Phenomenological Fact 3.11. Qualia can be recognised to a certain extend: Experiencing subjects can identify qualia which they have previously experienced.

Example 3.12. Concerning qualia of the ‘what is it like to be’ type, Phenomenological Fact 3.11 merely states that we can recognize whether what it is like to experience A (to taste a beer, say) and what it is like to experience B (to taste a bubblegum, say) are the same or not.

Phenomenal Fact 3.11 is important because it is the basis of the ability of an experiencing subject to introduce labels for his/her qualia, i.e. a name or reference for
non-communicable aspects of his/her experience. In what follows, we assume that labels are chosen such that different qualia are associated with different labels and, using Phenomenological Fact 3.11, that the same label is used to denote various occurrences of the same quale. Furthermore, we assume that all experiencing subjects use the same set of labels, which we denote by $L$. For our purposes, $L$ can be any set, which labels the set consists of does not matter in what follows.

The second phenomenological fact expresses the observation that something can be said about how non-communicable aspects occur in, or constitute, experience. In [Nag74], it corresponds to the observation that “structural features of perception might be more accessible to objective description, even though something would be left out” [Nag74 p. 449]. In [Cha96], it corresponds to the observation that “even if experiences are in some sense ‘ineffable,’ relations between experiences are not; we have no trouble discussing these relations, whether they be relations of similarity and difference, geometric relations, relations of intensity, and so on. As Schlick (1938) pointed out, the form of experience seems to be straightforwardly communicable, even if the content (intrinsic quality) is not.” [Cha96 p. 224]

**Phenomenological Fact 3.13.** Qualia have relations that can be communicated within the class $C$.

By Definition 3.4, this is a claim about experiences of all experiencing subjects in the class $C$. What is crucial is the communicability of the relations. It implies that we may represent the relations on the set of labels $L$ and assume (i.e. ask, cf. Footnote 12) labels to be chosen in such a way that the experienced relations between qualia are reflected in the relations represented on the labels.

**Example 3.14.** For qualia of the ‘what it is like to be’ type (introduced in Example 3.10) these relations include

- **Similarity.** Two qualia can be more or less similar.
- **Intensity.** A quale can occur in more or less intense versions.

among others.  

**Example 3.15.** Experiencing subjects typically experience some pairs of colours as similar to each other, whereas they experience others as not similar. E.g., small changes in hue usually result in colours which are perceived as similar, whereas large changes in hue result in colours which are not experienced as similar.

What is crucial for our purposes is that one may (and in practise often does) represent the experience of similarity of colours on the set of colour labels. Correspondingly, one may (and in practise often does) ask experiencing subjects to choose labels for their colour experiences.

---

12Note that throughout this section, assumptions are in fact conventions. E.g., this assumption can be satisfied by asking experiencing subjects (in an experiment, say) to choose labels as described. The assumptions can be made ‘without loss of generality’, so to speak.

13Similarity and intensity are simple examples of communicable relations between qualia. There may be many more communicable relations which express facts about how qualia appear in experience, some of which may only relate qualia of a particular type to each other. Further examples arguably include: Composition: Some qualia are experienced as a composition of two (or more) different qualia. I.e., the composed quale is but a combination (or simultaneous experience) of the composing qualia. Inclusion: Some quale may be experienced as containing one (or more) other quale. Here, the contained quale is but an aspect of the containing quale. Also, the distinction between various types (visual, auditory, tactile, etc.) of non-communicable aspects of experience is a relation in the sense of Phenomenological Fact 3.13.
in such a way that colours which are experienced as similar are similar according to the representation on colour space \(F\). We will study this in detail in Example 3.18 below.

Any representation of a relation on a set is mathematical in nature, whether written in terms of some formal language or not. Therefore, representations of communicable relations on the set \(L\) are mathematical in nature. They give rise to either a relation on \(L\) in the mathematical sense of the word (i.e. a subset \(R\) of \(L \times L\)) or to some more involved mathematical structure. Several detailed examples are given in Section 3.1.4 below. We refer to the set of labels \(L\) together with its mathematical structure as

\[
\text{label space } L. \tag{3.1}
\]

This space is the desired formal representation of experience. What is crucial about the introduction of labels is that one avoids from the very beginning an implicit assumption of an empirically well-defined method to refer to qualia of an experiencing subject. In any scientific investigation, whether related to model building or experiment, one can only refer to labels of qualia and has to take into account the ambiguity thereby induced, as we now explain.

### 3.1.2. References to Qualia

To describe the ambiguity of any reference to qualia mathematically, we proceed in two steps. First, we discuss the case where an experiencing subject uses labels to report on his/her experience without taking into account the communicable relations established in Phenomenal Fact 3.13. In this case, the experiencing subject is free to choose any label to denote any quale, the only requirements being that different labels are used for different qualia and that the same label is being used for a recurrent quale. We call a choice of labels for the qualia of an experiencing subject a labelling and use the term relabelling to denote a change of labelling. In the present case, a relabelling is simply a map

\[
s : L \to L, \quad \ell \mapsto s(\ell), \tag{3.2}
\]

which determines which label \(s(\ell)\) replaces the previous label \(\ell\). Since different qualia are required to carry different labels, this map is injective. Since it furthermore has domain and codomain \(L\), it is bijective. Since any composition of two relabellings of the form (3.2) yields another relabelling, and since due to the bijectivity, each map (3.2) is invertible, all possible relabellings form a group: The group of all bijective maps from \(L\) to itself. This group is called the symmetric group of the set \(L\).

The crucial insight here is that the group of relabellings allows to quantify the ambiguity of any statement that refers to labels of qualia. Consider e.g. the case where a statement only involves one label \(\ell_1 \in L\). Since we are disregarding communicable relations at this point, this statement could just as well have been formulated with any other label \(\ell_2 \in L\), simply because an experiencing subject may choose any label whatsoever to denote any quale. Mathematically, this is reflected by the fact that there is at least one relabelling \(s\) such that \(s(\ell_1) = \ell_2\). The same reasoning can be applied to sequences \((\ell_1, \ldots, \ell_n)\) of labels, e.g. obtained by verbal reports at subsequent times. The ambiguity of a sequence \((\ell_1, \ldots, \ell_n)\) of labels is the set of all sequences \((\ell'_1, \ldots, \ell'_n)\)

\(^{14}\)Note that this example is complicated by the fact that we calibrate colour experiences in practise: We apply or learn rules on how to pick colour labels related to external events such as wave-length impinging on the eye. This will be discussed in detail in Example 3.18 below. What is crucial is that a priori, individual labels so chosen do not correlate with colour experience: Two experiencing subjects may have a completely different colour experience despite using the same label ‘blue’. 

\[\hat{\text{\bf{\color{red}{(3.1)}\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.2)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.3)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.4)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.5)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.6)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.7)\quad}}}\]

\[\hat{\text{\bf{\color{red}{(3.8)\quad}}}\]
which can be obtained from the former by a relabelling $s$, i.e. the set of all sequences $(\ell_1', \ldots, \ell_n')$ for which there exists a relabelling $s$ such that $(\ell_1', \ldots, \ell_n') = (s(\ell_1), \ldots, s(\ell_n))$.

These statements are in fact statements about equivalence classes. To see this, define two labels $\ell_1$ and $\ell_2$ to be equivalent, $\ell_1 \sim \ell_2$, if and only if there exists a relabelling $s$ such that $s(\ell_1) = \ell_2$. The ambiguity of a label $\ell \in L$ is given precisely by the equivalence class of this label,

$$[\ell] := \{ \ell' \mid \ell' \sim \ell \} = \{ \ell' \mid \exists s : \ell' = s(\ell) \},$$

(3.3)

because this class contains all labels which an experiencing subject could have chosen. The same is true for sequences: If we define two sequences to be equivalent, $(\ell_1', \ldots, \ell_n') \sim (\ell_1, \ldots, \ell_n)$, if and only if there exists a relabelling $s$ such that $(\ell_1', \ldots, \ell_n') = (s(\ell_1), \ldots, s(\ell_n))$, the ambiguity of a sequence of labels is given precisely by the equivalence class of this sequence,

$$[(\ell_1, \ldots, \ell_n)] := \{ (\ell_1', \ldots, \ell_n') \mid (\ell_1', \ldots, \ell_n') \sim (\ell_1, \ldots, \ell_n) \},$$

(3.4)

because this class contains precisely all those descriptions of the sequence which an experiencing subject may give. Another way to put this is that the equivalence classes (3.3) and (3.4) are what is empirically well-defined, not the labels themselves, these only have meaning for the experiencing subject him/herself once he/she has chosen a particular labelling.

Next, we take into account the communicable relations between qualia established in Phenomenological Fact 3.13. Recall that a phenomenological fact is defined to be a statement about aspects of experience which holds throughout all experiences of the class $\mathcal{C}$ of experiencing subjects. Since we have chosen $\mathcal{C}$ to comprise all adult humans, Phenomenological Fact 3.13 is a claim that there are relations between qualia which all adult humans experience, and that they furthermore can collate the fact that they have experienced qualia as so related. In simple terms, Phenomenological Fact 3.13 thus expresses the fact that something can be said about non-communicable aspects of experience, something about how they appear in experience. This fact makes the ambiguity of labels much smaller.\(^{15}\)

To quantify this, we work with the label space $L$ introduced above: I.e., we assume that the relations between qualia have been represented on the set of labels\(^{16}\) and ask experiencing subjects to pick labels for the qualia they experience in accordance with this representation. As above, we refer to any such choice as labelling. This implies that the freedom of every experiencing subject to choose labels is smaller than in the case above: Functions (3.2) only constitute relabellings if they preserve the communicable relations represented on the set $L$, i.e. if they preserve the structure of the space $L$. A bijective function from a space to itself which preserves the structure of this space is called an automorphism of the space. Thus in

\(^{15}\)In other words, the symmetric group of the set $L$ does not capture the ambiguity of qualia we find in experience. This is particularly apparent from the fact that the symmetric group allows to map any choice $(\ell_1, \ldots, \ell_n)$ of labels to any other choice $(\ell_1', \ldots, \ell_n')$, provided that every label occurs at most once in each choice.

\(^{16}\)For all practical purposes, one can obtain such a representation by simply asking one experiencing subject to pick a labelling and to report, in terms of this labelling, on his/her experienced relations. Other experiencing subjects are then required to choose labels according to this representation. For details, see Example 3.13 below. For explicit examples on how such a representation might look, cf. Examples 3.19 to 3.22 below.
the case where we take into account the communicable relations of qualia, *relabellings* are elements of the

\[
\text{automorphism group } \text{Aut}(L).
\]

(3.5)

We summarize this by saying that the automorphism group \( \text{Aut}(L) \) describes the *freedom of relabelling* of every experiencing subject. We can now argue exactly as above to identify the ambiguity of any statement that uses a sequence \((\ell_1, \ldots, \ell_n)\) of labels. The result is that the ambiguity is given precisely by the equivalence class

\[
[(\ell_1, \ldots, \ell_n)] := \{ (\ell'_1, \ldots, \ell'_n) \mid (\ell'_1, \ldots, \ell'_n) \sim (\ell_1, \ldots, \ell_n) \},
\]

(3.6)

where \( \sim \) denotes the equivalence relation defined as

\[
(\ell'_1, \ldots, \ell'_n) \sim (\ell_1, \ldots, \ell_n) \quad \text{if and only if there is an } s \in \text{Aut}(L) \text{ such that } \ell'_i = s(\ell_i) \text{ for all } i = 1, \ldots, n.
\]

(3.7)

This class contains precisely all descriptions of the sequence of qualia an experiencing subject might give: A description in every possible labelling. To obtain the ambiguity of individual labels, we simply set \( n = 1. \)

**Remark 3.16.** In practise, we typically establish labels by reference to particular “external” events, such as particular wavelengths emerging from a light source in the case of colour experiences. Socially established labels of this sort are of course very useful in various circumstances, precisely because they correlate with external events. However, a priori there is no reason to assume that qualia of different experiencing subjects which are denoted by the same label are the same, even if the labels correlate with the same external event. In fact, an assumption of this kind has no empirical meaning because the definition of qualia implies that neither the identity of qualia of different experiencing subjects with an external event, nor the equality of qualia of different experiencing subjects can be empirically tested. Statements of this sort can only be meaningful if formulated based on a scientific methodology which is compatible with the non-communicability of the aspects of experience under consideration.

3.1.3. Summary. In summary, we have seen that the empirically well-defined references to qualia are elements of the quotient space

\[
L^{\times n} / \sim,
\]

(3.8)

where \( L \) is the space \( \overset{\text{[3.1]}}{\text{L}} \) whose structure represents communicable relations between qualia, where \( \sim \) denotes the equivalence relation \( \overset{\text{[3.7]}}{\text{3.7}} \) and where \( n \in \mathbb{N} \) is the length of a sequence. This quotient space describes what, concerning non-communicable aspects of experience, is amenable to the usual scientific methodology, e.g. the one in use in neuroscience. If any of the equivalence classes \( \overset{\text{[3.6]}}{\text{3.6}} \) is non-empty, there is a question about experience which is not amenable to the usual methodology: Why one member of the equivalence class, rather than another, has been experienced. The essential purpose of Section \( \overset{\text{[3.4]}}{\text{3.4}} \) is to show that this question is amenable to a scientific inquiry, if the latter is suitably defined. The goal of Section \( \overset{\text{[5]}}{\text{5}} \) is to show that under certain circumstances, this seemingly minor question may have large empirical consequences.

Coming back to Definition \( \overset{\text{[1.1]}}{\text{1.1}} \) we summarize what has been said as follows:

\[\text{\[\overset{\text{[17]}}{\text{17}}\]}\] One may even take this to be unlikely, given the difference of neural structure across individuals.
Definition 3.17. What is to be studied by the scientific study of consciousness according to the phenomenological grounding is experience and its relation to the physical domain. This includes the study of the quotient space (3.8), using the usual scientific methodology, but also the study of qualia proper, i.e. the study of those aspects of experience whose non-communicability prevents the standard methodology from being applied.

3.1.4. Examples. We close this section with several examples. First, in Example 3.18 we continue the discussion of colour experience and show that colour spaces, which are largely in use in commercial applications, constitute the label spaces for colour qualia as defined above. In Example 3.19 to 3.22 we consider possible mathematical structures of the label space $L$, some of which have been proposed in the literature.

Example 3.18. To illustrate the meaning of the label space $L$ and the group $\text{Aut}(L)$, as well as Remark 3.10 we consider again colour experiences. As we have explained in Example 3.7, these satisfy the defining property of qualia. We will generally denote the quale ‘what it is like to see light of wavelength $\lambda$’ as ‘experience of $\lambda$’. Furthermore, we will use the symbol $\bar{\lambda}$ to denote a mixture of light of varying wavelength.

We start by fixing a particular human observer, the “standard observer” [Kue10], and choose a set $L_{cl}$ that is in one-to-one correspondence to all colours which this human can experience. As usual in colour science, we assume that there is a large class $C$ of humans which have the same set of possible colour experiences as the standard observer. This assumption implies that every human in the class $C$ can specify a one-to-one correspondence between the set $L_{cl}$ and his/her colour experiences. The set $L_{cl}$ is thus a set of labels of colour qualia as introduced after Phenomenological Fact 3.11. It is also the basis of the definition of colour spaces (cf. below).

The set $L_{cl}$ can be calibrated: Since colour experiences arguably arise as a response to mixtures $\bar{\lambda}$ of light impending on the retina, we may identify every element $\ell \in L_{cl}$ with a particular mixture $\bar{\lambda}$. The set of mixtures visible to the human eye can, in turn, be represented as a subset $S \subset \mathbb{R}^3$, roughly speaking by taking the three components of a vector $v \in S$ to represent the relative intensities of three reference wavelengths. Putting these two steps together, we may in fact choose the set $L_{cl}$ to be the subset $S \subset \mathbb{R}^3$. In this case, every label $\ell \in L_{cl}$ is a 3-tuple of real numbers which specifies which mixture $\bar{\lambda}$ of light has to be presented to a particular human to evoke the quale that he/she has denoted by that very label $\ell$.

This calibration may lead one to think that there is a unique way of referring to colour qualia. However, this is not the case. To see this, assume that we fix some label/vector $\ell \in S = L_{cl}$ as well as two experiencing subjects $A$ and $B$. Let us denote the mixture of light that corresponds to this vector as $\bar{\lambda}_\ell$. When we present this mixture $\bar{\lambda}_\ell$ to the two experiencing subjects, subject $A$ has the colour experience he/she has labelled as $\ell$, and so does subject $B$. However, this has nothing to say on whether the colour experiences are the same or not: E.g., subject $B$ might have the colour experience subject $A$ is having upon presentation of a completely different mixture $\bar{\lambda}_w \neq \bar{\lambda}_\ell$.

This illustrates the fundamental difficulty related to qualia as defined in Definition 3.8. If we would “know” (e.g. as the result of some scientific investigation) that the presentation of the same colour stimuli $\bar{\lambda}_w$ to various subjects results in them having the same colour experience, we could meaningfully talk, or refer to, colour experiences of different subjects in terms of...
stimuli. More generally, if statements of the type

\[ \text{"subject A will have colour experience } X_1 \text{ once presented input } \lambda \text{"} \tag{3.9} \]

would be known, these statements would allow us to directly refer to A’s colour experiences, putting us into the position to do science as usual. However, the fundamental difficulty of the subject is that statements like (3.9) do not carry any intersubjective meaning at all: Due to the impossibility of communicating colour experiences, statement (3.9) cannot be distinguished (by anyone but subject A) from the statement

\[ \text{"subject A will have colour experience } X_2 \text{ once presented input } \lambda \text{"}, \]

where \( X_2 \) is any colour experience of A with the same unary communicable relations (such as intensity). This problem exists independently of whether we consider the statement (3.9) to be a hypothesis or to be the result of some purported scientific investigation. Statements of this type do not have unambiguous intersubjective meaning.

As explained above, what has intersubjective meaning are the equivalence classes (3.6). They express facts about colour experience which are invariant with respect to the labelling that an experiencing subject chooses. We now illustrate this in detail for colour qualia.

First, we need to find the communicable relations between colour experiences referred to in Phenomenological Fact 3.13. Luckily, this has been on the agenda of colour science for decades, so that we may simply turn to its results. Put in simple terms, there seem to be three types of communicable relations [Kue10]: Continuity of change of colours (whether some time-continuous sequence of colour experiences is perceived as continuous or not), behaviour under mixtures of colours (whether a mixture of two colour experiences is perceived as equal to another colour experience or not) and (less well known) a notion of distance of colours (whether two colour experiences are perceived as more different to each other than another pair of colour experiences).

Next, we need to translate these communicable relations into mathematical structures on the set \( L_{cl} \). This yields the label space (3.1) of colour qualia. Again, colour scientists have done the work for us: They have defined colour spaces in order to formalize these communicable relations [Kue10]. A colour space is a closed subset \( S \) of \( \mathbb{R}^3 \) which is in a one-to-one correspondence with all colours humans may experience, chosen such that continuity is represented by the induced topology of \( \mathbb{R}^3 \) (a path of colours experiences is continuous if the labels form a continuous path in the colour space), mixture is represented by straight lines (equal mixing of two colour experiences \( \ell_1 \) and \( \ell_2 \) yields the colour experience that carries the label that is at the center of the straight line that connects \( \ell_1 \) and \( \ell_2 \)), and finally experienced distance of colour qualia is represented by a metric on \( S \). Thus a colour space is a label space (3.1) for colour qualia.

There are many subsets of \( \mathbb{R}^3 \) which satisfy these requirements: For any choice of subset \( S \), there is a large class of transformation of \( \mathbb{R}^3 \) which, together with a corresponding transformation of the metric, yield another subset \( S' \) of \( \mathbb{R}^3 \) which equally represents colour experiences as well as their communicable relation. Colour science uses the calibration described above to fix specific choices of subsets \( S \), so that the coordinates of the elements of \( S \) can be translated into mixtures of wavelengths \( \lambda \). However, as explained above, for the study of colour experience, calibrations do not have any relevance a priori, so that no particular choice of subset can be singled out.

In order to specify the group of relabellings for this example, we note that in more abstract terms, a colour space is a smooth 3-dimensional Riemannian manifold: Its topology represents the continuous changes of colour experience and its metric \( g \) specifies both the geodesics (generalized “straight lines”), which describe the mixture of colour experiences, as well as a

\[ \text{Cf. [Kue10]. However, note that a more axiomatic treatment may result in different mathematical spaces [Res74, Pro17]. Furthermore, the assumption of smoothness may not be justified and one might have to consider manifolds with corners.} \]
distance function which describes the experience of distance between colour qualia. The various choices of subsets $S$ of $\mathbb{R}^3$ correspond precisely to choices of coordinates of this manifold. We summarize this as
\[ L = (L_{cl}, g). \] (3.10)
This is the actual form of the label space $[3.1]$ of colour qualia. Its elements label the set of colour experiences and its structure represents the communicable relations between them. An experiencing subject can specify his/her colour experiences by specifying points (in the case of individual colour experiences) or curves (in the case of time-continuous colour experiences) on this manifold. The freedom of choosing labels is described by the automorphism group of $L$. In the case $[3.10]$ of a Riemannian manifold, this is the group of isometries, i.e. diffeomorphisms which leave the metric invariant:
\[ \text{Aut}(L) = \text{Iso}(L). \]
Thus the ambiguity of any statement in terms of colour labels $(\ell_1, ... , \ell_n)$ is given by the equivalence class
\[ [(\ell_1, ... , \ell_n)] \]
which is defined as in $[3.6]$ with two sequences being equivalent if there is an isometry $s \in \text{Iso}(L)$ which transforms every element of the first sequence into the corresponding element of the second sequence. The actual form of the equivalence classes depends on the metric $g$, which can be determined experimentally. The current version of the distance function internationally in use reviewed e.g. in [SWD04], a discussion of which however goes beyond the scope of this example.

Putting everything together, we conclude that any statement, scientific or otherwise, that addresses colour experiences sensu stricto – i.e. which addresses what it is like to experience colours – only makes sense if it is invariant with respect to $\text{Iso}(L)$ transformations. This is a consequence of the fact that qualia are non-communicable and of the corresponding freedom of every experiencing subject to choose names for the qualia he/she experiences.

The difference between labels of colour experiences and colour experiences, though arguably not much of importance in daily life, can be crucial for scientific investigations. For example, if a study compares the calibrated label $\ell$ that a subject reports with neural activity, it does not investigate the relation between neural activity and colour experience but rather the relation between neural activity and presentation of wavelengths $\lambda$ to the retina. These two objects of investigation refer to completely different scientific agendas.

Example 3.19. Pretopological structure on $L$. For this example, we consider the similarity relation between qualia introduced in Phenomenological Fact 3.13 but understood in a binary way (two qualia ‘are similar’ or are ‘not similar’, no varying degrees of similarity). This similarity relation can be used to define a pretopological structure on $L$, as explained in detail in [Pre19] in a slightly different conceptual setting. We now summarize this construction:

\[ \text{Aut}(L) = \text{Iso}(L). \]

Thus the ambiguity of any statement in terms of colour labels $(\ell_1, ... , \ell_n)$ is given by the equivalence class
\[ [(\ell_1, ... , \ell_n)] \]
which is defined as in [3.6] with two sequences being equivalent if there is an isometry $s \in \text{Iso}(L)$ which transforms every element of the first sequence into the corresponding element of the second sequence. The actual form of the equivalence classes depends on the metric $g$, which can be determined experimentally. The current version of the distance function internationally in use reviewed e.g. in [SWD04], a discussion of which however goes beyond the scope of this example.

Putting everything together, we conclude that any statement, scientific or otherwise, that addresses colour experiences sensu stricto – i.e. which addresses what it is like to experience colours – only makes sense if it is invariant with respect to $\text{Iso}(L)$ transformations. This is a consequence of the fact that qualia are non-communicable and of the corresponding freedom of every experiencing subject to choose names for the qualia he/she experiences.

The difference between labels of colour experiences and colour experiences, though arguably not much of importance in daily life, can be crucial for scientific investigations. For example, if a study compares the calibrated label $\ell$ that a subject reports with neural activity, it does not investigate the relation between neural activity and colour experience but rather the relation between neural activity and presentation of wavelengths $\lambda$ to the retina. These two objects of investigation refer to completely different scientific agendas.

Example 3.19. Pretopological structure on $L$. For this example, we consider the similarity relation between qualia introduced in Phenomenological Fact 3.13 but understood in a binary way (two qualia ‘are similar’ or are ‘not similar’, no varying degrees of similarity). This similarity relation can be used to define a pretopological structure on $L$, as explained in detail in [Pre19] in a slightly different conceptual setting. We now summarize this construction:

\[ \text{Aut}(L) = \text{Iso}(L). \]

Since the ordering of distances between pairs of colours, rather than the numerical value of the distance itself, is communicable, one could make the point that the relabelling freedom is given by the group of diffeomorphisms which leave the metric invariant up to a conformal factor. Since the present example is, mainly, of a pedagogical interest, we do not explore this further at this point. Cf. also Footnote 19.

We note that it is possible that some sequences $(\ell_1, ..., \ell_n)$ are not ambiguous, i.e. that
\[ [(\ell_1, ..., \ell_n)] = \{(\ell_1, ..., \ell_n)\}. \]
This means that there is one unique sequence of colour experiences which has the properties represented by the sequence $(\ell_1, ..., \ell_n)$ of labels, or put differently, that there is only one possible choice of labels for this sequence that takes into account the communicable relations as described. Sequences of this kind may be used to remove the ambiguity of the labels they contain and make these aspects of experience accessible to a proper scientific analysis.

The following definitions and their relation to topology are intuitively accessible if one thinks about open balls in a metric space such as $\mathbb{R}^3$, where $\circ$ is defined as overlap.
First, we define a binary relation $R_\subseteq \subset L \times L$ on $L$. If two qualia with labels $\ell_1$ and $\ell_2$ are perceived as similar by an experiencing subject, we define the corresponding labels to be related according to $R_\subseteq$, which we denote as $\ell_1 \sim \ell_2$ (i.e., $\ell_1 \sim \ell_2 \iff (\ell_1, \ell_2) \in R_\subseteq$, and similarly below). Thus $R_\subseteq$ is given directly by experience. We assume that $\ell \sim \ell$ for all $\ell \in L$. Second, based on the data of $R_\subseteq$, we define another relation $R_\ll$ on $L$, called “parthood relation” [Pre19] as

$$\ell_1 \leq \ell_2 \iff \ell \circ \ell_1 \Rightarrow \ell \circ \ell_2.$$ 

Thus $\ell_1 \leq \ell_2$ holds iff all qualia which are similar to $\ell_1$ are also similar to $\ell_2$. Third, we use the parthood relation $R_\ll$ to define yet another relation $R_\ll\sim$, called “connection”, as follows:

$$\ell_1 \sim \ell_2 \quad \text{iff} \quad \exists \tilde{\ell} \in L \text{ such that } \tilde{\ell} \circ \ell_1 \text{ and } \tilde{\ell} \circ \ell_2$$

as well as $\ell \leq \tilde{\ell} \Rightarrow \ell \circ \ell_1 \text{ or } \ell \circ \ell_2$.

Note that $\ell_1 \leq \ell_2$ implies $\ell_1 \sim \ell_2$. We extend this notation to sets $A \subset L$ by defining

$$\ell \sim A \quad \text{iff} \quad \ell \sim \tilde{\ell} \text{ for at least one } \tilde{\ell} \in A.$$ 

This allows us to define an operator $pcl$, which takes a subset $A \subset L$ to another subset $pcl(A)$ which contains all qualia which are connected to at least one of the qualia in $A$:

$$pcl(A) := \{\ell \mid \ell \sim A\}.$$ 

The operator $pcl$ satisfies three of the four Kuratowski closure axioms [Per64, Sec. 3.2], but may not satisfy $pcl(pcl(A)) = pcl(A)$ for all $A \subset L$ (idempotence). Hence it constitutes a preclosure operator, so that $(L, pcl)$ constitute a pretopological space [nLa19].

In order to define what constitutes a relabelling in this example, we note that a function $f$ between two pretopological spaces $(L, pcl)$ and $(L', pcl')$ is defined to be continuous if

$$f(pcl(A)) \subseteq pcl'(f(A))$$

for all $A \subset L$. The automorphism group $Aut(L)$ of $(L, pcl)$ is the set of all continuous invertible functions $f : L \rightarrow L$ whose inverse is also continuous, with group operation given by function composition.

This example is important because it shows neatly how non-trivial mathematical properties of the label space $L$ can be defined directly in terms of experienced relations between qualia. The similarity relation established via Phenomenological Fact 3.13 may, of course, not actually be binary: There seem to be various degrees, maybe even a continuum, of similarities of qualia.

\textbf{Example 3.20. Partial order on $L$.} Our next example goes back to [Res18]. First, we observe that next to the two relations mentioned in Phenomenological Fact 3.13, qualia may in fact have compositional relations that can be communicated: An experiencing subject may find that the ineffable aspect of an experience he/she is having at a particular time includes an ineffable aspect he/she has had at another time. In this case, we may say that the former quale includes the latter quale. If $\ell_1$ is the label which the experiencing subject has chosen for the former quale and $\ell_2$ is the label he/she has chosen for the latter quale, we will denote this relation between the two qualia as $\ell_2 \leq \ell_1$.

By convention, we may put $\ell \leq \ell$ for all $\ell \in L$ (reflexivity). Furthermore, it is reasonable to hold that if both $\ell_1 \leq \ell_2$ and $\ell_2 \leq \ell_1$ for two labels $\ell_1, \ell_2 \in L$, these labels actually refer to the same quale, so that $\ell_1 = \ell_2$ (anti-symmetry). Finally, qualia seem to satisfy that $\ell_1 \leq \ell_2$ and $\ell_2 \leq \ell_3$ imply $\ell_1 \leq \ell_3$ (transitivity). Therefore, this actually constitutes a partial order on $L$ and turns $(L, \leq)$ into a partially ordered set. The automorphism group consists of bijective functions $f : L \rightarrow L$ which are order-embedding, i.e. which satisfy $\ell_1 \leq \ell_2$ if and only if $f(\ell_1) \leq f(\ell_2)$ for all $\ell_1, \ell_2 \in L$. Thus one can see nicely that the automorphism group describes the freedom of relabelling: Its elements represent changes of labels which preserve the inclusion relation between qualia. \hfill \lozenge
Example 3.21. **Involutive semigroup structure on** $L$. This example also goes back to [Res18]. In order to state it, note that in Definition 3.1, we have defined the term ‘experience’ with respect to instants of time. This implies that qualia (being aspects of experience) are associated to a instant of time as well \(^{23}\), therefore excluding a sequence of two qualia arising at two consecutive instants of time to constitute another quale. However, one might drop this restriction to instants of time, and define qualia as aspects of experience in general. Following this line of thought, one could argue that for any two qualia $\ell_1, \ell_2$, there is another quale $\ell_3$ which is the consecutive experience of the two qualia. One might denote $\ell_3$ as

$$\ell_3 = \ell_1 \& \ell_2,$$

where the ‘$\&$’ represents “and then” [Res18]. If one furthermore demands associativity, which does seem to be plausible, this defines a semigroup $(L, \&)$.

Next, one may consider an operation which reverses this temporal order of qualia. This may or may not have deep conceptual meaning: On the one hand, it may merely map any quale of the form $\ell_1 \& \ell_2$ to a quale of the form $\ell_2 \& \ell_1$, both of which have to exist due to the semigroup structure introduced above. On the other hand, it may express a deep fact about reversal of psychological time [Res18]. In both cases, skipping over a few technical details, this gives rise to an involution [Res18], i.e. a map

$$* : L \rightarrow L \text{ such that } \ell \mapsto \ell^* \quad (\ell^*)^* = \ell.$$  

In summary, the time composition relation of qualia may be represented on the space of labels in terms of an involutive semigroup structure.

Example 3.22. **Hilbert space structure on** $L$. The last example is intended to evaluate in how far the axioms of a Hilbert space can be grounded in the relations introduced in Phenomenal Fact 3.13. The upshot is that whereas some of the axioms can be motivated based on Phenomenological Fact 3.13, others cannot. In what follows, we make several assumptions about the set of all experiences which an experiencing subject might have. These assumptions are phenomenological in flavour, yet some may ultimately not be justified.

(A1) We assume that with respect to any two qualia of one experiencing subject, the experiencing subject might have an experience which has exactly these qualia as ineffable aspects.

With respect to qualia of the ‘what it is like to be’ type (Example 3.10), this assumption amounts to the following statement: If an experiencing subject has made an experience which included an ineffable ‘what is it like to be’ aspect (quale) which he/she labels by $\ell_1$, and another experience which included an ineffable ‘what is it like to be’ aspect (quale) which he/she labels by $\ell_2$, then it is possible that he/she will make an experience which has exactly $\ell_1$ and $\ell_2$ as ineffable aspects. We will use the term ‘simultaneous experience of $\ell_1$ and $\ell_2$’ as an abbreviation for the statement that the experiencing subject in question has an experience which includes both aspects $\ell_1$ and $\ell_2$. To give an example, let $\ell_1$ refer to what it is like to taste cheese and $\ell_2$ refer to what it is like to smell wine. In this case, Assumption (A1) amounts to granting the possibility of the experiencing subject in question simultaneously experiencing what it is like to taste cheese and what it is like to smell wine. Whether this experience actually arises when the subject eats cheese and drinks wine is of no concern with respect to Assumption (A1). We take the combination of the same experience $\ell$ with itself as denoting the experience of quale $\ell$ but twice as intense (cf. below). In order to motivate a group structure with respect to simultaneous experience, the following assumption is necessary:

---

\(^{23}\)The term ‘instant of time’ may refer to experiential instants of time or to instants of time as used in physics, i.e. points $t \in \mathbb{R}$. 

(A2) We assume that there is a unique neutral quale which we denote by \( \mathcal{0} \). Furthermore, we assume that for every quale \( \ell \), there is an quale \(-\ell\) such that an experience which includes both \( \ell \) and \(-\ell\) is not distinguishable from (and hence equal to) the experience of the neutral quale.

It seems that this assumption is utterly beyond empirical justification, since it invokes something like “cancellation” of ‘what is it like to be’ aspects of experiences, so that we may only be able to ground a semigroup-structure of qualia with respect to combination (‘simultaneous experience’). For the purpose of this example, we proceed nevertheless. We denote the simultaneous experience of two qualia \( \ell_1 \) and \( \ell_2 \) by \( \oplus \), so that the ineffable aspect of the experience which comprises both qualia labelled as \( \ell_1 \) and as \( \ell_2 \) established by Assumption (A1) is labelled by \( \ell_1 \oplus \ell_2 \). Associativity and commutativity hold, so that we have:

- (A1) and (A2) imply that \( \oplus : L \times L \to L \) is an abelian group.

Next, we model changes of intensity, as conceded in Phenomenological Fact 3.13 by a positive real number in the following sense: If \( \ell_2 \) is the same quale as \( \ell_1 \), but \( c \) times more intense, then we denote \( \ell_2 = c\ell_1 \), where \( c \in \mathbb{R}^+ \). For \( c \in \mathbb{R}^- \), \( c\ell_1 \) is the opposite experience \(-\ell_1 \) introduced in (A2), but experienced \(|c|\) times as intense as \(-\ell_1 \), where \(|c|\) is the modulus of \( c \).

Finally, we assume that as intensity decreases, \( c \to 0 \), any experience goes over to the neutral quale, formally \( \lim_{c \to 0} c\ell = \mathcal{0} \) for any \( \ell \in L \), where \( \mathcal{0} \) denotes the neutral quale introduced in (A2). Making the idealized assumption that a continuum of more and less intense versions of any experience is possible, we have:

- The intensity relation of Phenomenal Fact 3.13 may be taken to give rise to a scalar multiplication \( \odot : \mathbb{R} \times L \to L \).

As usual, we suppress the symbol \( \odot \) for scalar multiplication. We need to check whether the axioms of a vector space relating scalar multiplication and addition hold. Our interpretation implies that \( 1\ell = \ell \), hence neutrality of \( 1 \in \mathbb{R} \) holds. The two axioms of distributivity read

\[
\begin{align*}
(c \odot (\ell \oplus \ell')) &= (c\ell) \oplus (c\ell') \quad \text{(3.11)} \\
((c + c')\ell) &= c\ell \odot c'\ell \quad \text{for all } c, c' \in \mathbb{R} \text{ and } \ell, \ell' \in L \quad \text{(3.12)}
\end{align*}
\]

Axiom 3.11 says that a \( c \) times more intense simultaneous experience of \( \ell \) and \( \ell' \) arises as the combination of \( c \) times more intense experiences of \( \ell \) and \( \ell' \), respectively, which we take as a plausible assumption in the context of this example. Axiom 3.12 states a compatibility of addition of intensities with combinations of experience. E.g., it says that an experience \( \ell' \) which is the same as another experience \( \ell \) but twice as intense, \( \ell' = 2\ell \) can arise as the simultaneous experience of the combination of \( \ell \) with itself. We render this axiom at least somewhat plausible by defining the combination of an experience with itself to be the same experience experienced twice as intense. Finally, we note that the associativity axiom \( (cc')\ell = c(c'\ell) \) is compatible with our interpretation of \( \odot \) and \( \oplus \). We therefore have:

- \((L, \oplus, \odot)\) satisfies the axioms of a vector space.

It remains to implement the the relation of similarity between qualia. As before, we idealize and assume that there is a non-negative real number which specifies how similar two qualia \( \ell_1 \) and \( \ell_2 \) are. We denote this number by \( \langle \ell_1, \ell_2 \rangle \). If \( \ell_1, \ell_2 \) are not similar at all, we set \( \langle \ell_1, \ell_2 \rangle = 0 \). If they are similar to some degree, we have \( \langle \ell_1, \ell_2 \rangle > 0 \), where a larger value implies more similarity. It seems natural to impose symmetry, \( \langle \ell, \ell' \rangle = \langle \ell', \ell \rangle \) for all \( \ell, \ell' \in L \).

An inner product furthermore satisfies

\[
\begin{align*}
\langle \ell, \ell \rangle &= 0 \iff \ell = 0 \quad \text{(Definiteness)} \\
\langle \ell, c\ell' \rangle &= c \langle \ell, \ell' \rangle \quad \text{(Linearity)} \\
\langle \ell, \ell' \oplus \ell'' \rangle &= \langle \ell, \ell' \rangle + \langle \ell, \ell'' \rangle
\end{align*}
\]

for all \( c \in \mathbb{R} \) and \( \ell, \ell', \ell'' \in L \). Out of those three axioms, only the last one seems reasonable to some extent. It says that similarity is compatible with simultaneous experience: The similarity
between a quale ℓ and the simultaneous experience of qualia ℓ′ and ℓ″ is given by the sum of the similarity of the quale ℓ to each one of the qualia ℓ′ and ℓ″.

Definiteness says that the only quale which is not similar to itself is the neutral quale. This seems rather problematic if one chooses the interpretation of 0 introduced in (A2). The first axiom of linearity says that the similarity between a quale ℓ and a c times more intense version of a quale ℓ′ is given by c times the similarity between ℓ and ℓ′. As mentioned before, in order to have a nice and clear example, we will accept also these assumptions for now, so that in summary we have:

- (L, ⊕, ⊙, ⟨, ⟩) satisfies the axioms of a inner product space or Pre-Hilbert space.

The inner product ⟨, ⟩ introduces a norm on L as usual by ∥ℓ∥ = √⟨ℓ, ℓ⟩. This norm may be interpreted as the intensity of a quale ℓ.

The inner product space (L, ⊕, ⊙, ⟨, ⟩) may not be complete with respect to this norm, meaning that there are Cauchy sequences in L which do not converge to an element in L. In terms of qualia, this means that there are sequences of qualia whose elements become ever more similar to each other but which do not converge to any quale in the topology specified by the similarity relation. In order to exclude such cases, we consider the completion of L with respect to the norm ∥ ∥, which is unique up to isometric isomorphism. Alternatively, we may assume that there is a finite number of classes of non-similar qualia, so that completeness holds automatically. A complete inner product space is a Hilbert space. Denoting, as usual, completion by a line over the corresponding quantities, we have:

- The label space L carries the structure of a real Hilbert space (L, ⊕, ⊙, ⟨, ⟩), which we denote by H_L.

Note that this is an abstract Hilbert space: Due to the ineffability of qualia, the elements of the Hilbert space do not have an intrinsic communicable nature (as e.g. the case if one considers function spaces). The automorphism group Aut(L) is the group U(H_L) of unitary operators.

3.2. Explanatory Gap. An “explanatory gap” [Lev83] between a phenomenon and natural science occurs if the phenomenon has properties which render it incompatible with all notions of explanation used in natural science. This is in particular the case if the phenomenon violates a necessary condition for the application of any notion of explanation used in natural science.

Explanatory gaps are taken by some to indicate or entail ontological gaps (cf. [Cha10, Ch. 5, Sec. 3.4]), e.g. that something related to consciousness pertains to reality in addition to the physical domain. Whether this is legitimate or not is a philosophical question which we will not address here. What matters for us is that if there is an explanatory gap, a change of methodology is necessary if the phenomenon is to be addressed by scientific means. This change may or may not be motivated by ontological considerations.

In the context of qualia as specified in Definition 3.8, an explanatory gap arises if one accepts the claim that any explanation used throughout natural science presupposes that the phenomenon to be explained can be communicated. The justification of this claim will ultimately necessitate detailed investigations based on philosophy of

---

24Here, by ‘phenomenon’, we mean anything that occurs or manifests itself in a general sense, including both scientifically observable “empirical phenomena” (such as data of an experiment) as well as what is directly or indirectly perceived (experiences).

25This is not to say, of course, that every phenomenon can be addressed by scientific means. There may be phenomena to which the scientific method cannot be applied. However, it seems that the only way to establish whether this is the case for a particular phenomenon is to try to develop a suitable methodology and, if successful, to apply it.
explanation. As far as the review in [Woo17] goes, this claim seems to hold true for the
deductive-nomological model of explanation, the deductive-statistical model of expla-
nation, the statistical relevance model of explanation (where the explanandum needs
to have communicable attributes) and the causal mechanical model of explanation,
which is why we take it to be true for what follows.

Put in simple words, this explanatory gap says that there is no possibility at present
in natural science to explain why an experiencing subject experiences a particular
quale. Any purported explanation is unscientific insofar it does not have empirical
meaning. This, of course, does not hold for relations between qualia as established in
Phenomenological Fact 3.13, which can be described intersubjectively and hence may
be explained by contemporary natural science.

The purpose of Section 3.4 is to present a methodology which allows to study qualia
as defined in Definition 3.8 and their relation to the physical domain by scientific
means. This methodology is a generalization of Chalmers’ strategy outlined in Sec-
tion 2. In preparation, we recall some general facts about what constitutes a theory
in natural science.

3.3. Theories, Models, Symmetries. There are various different accounts in phi-
losophy of science of what constitutes a scientific theory. Roughly, one may distinguish
syntactic accounts, semantic accounts and pragmatic accounts [Win16], which differ
mainly in the role they attribute to mathematical formalization. Which account of
scientific theories is most adequate for a young field such as the scientific study of
consciousness is yet to be seen.

In order to make progress, we nevertheless need to commit to a small set of in-
gredients of a scientific theory which we can refer to in what follows. The following
choices are motivated by contemporary physical theories, but seem general enough to
not directly exclude any of the above-mentioned accounts of what constitutes a scien-
tific theory. Needless to say, this list is not intended to be a sufficient account of what
constitutes a scientific theory.

We assume that a theory $T$ includes:

- A set of dynamical variables $d$. (Those quantities whose variation is determined
  by $T$ to some extend.)

- Some background structure $b$. (Variables, or general mathematical structures,
  whose change is not determined by $T$. Background structure needs to be fixed
  in order to determine the variation of $d$ in a particular application.)

We assume that the variation or change of the dynamical variables $d$ can be expressed
with respect to some parameter $t$ which takes values in some set $\mathcal{I}$. Typically, the pa-
rameter is assumed continuous and interpreted as time. However, this is not necessary
in our case: The set $\mathcal{I}$ may or may not carry some mathematical structure (such as a
topology) and it may or may not be interpretable as time. E.g., $\mathcal{I}$ could just be a finite
set so as to give rise to a discrete parameter $t$. A family $(d_t)_{t \in \mathcal{I}}$ is a function $f : \mathcal{I} \rightarrow d$,
which we will call “trajectory”. It describes the change of dynamical variables with
respect to the parameter $t$. Continuing our list, a theory $T$ thus includes:

- A set of kinematically possible trajectories $K$. Sometimes, this includes all
  possible trajectories, $K = \{(d_t)_{t \in \mathcal{I}}\}$, but in many cases, trajectories need to

\[^{26}\text{We use the word ‘variable’ in a general sense here: A variable may represent something as simple}
\text{as a natural number just as well as an operator-valued field on some manifold.}\]
satisfy certain mathematical requirements, such as differentiability with respect to the parameter \(t\).

- Some laws \(\mathcal{L}\). (Typically equations or variational principles, but \(\mathcal{L}\) may also include different formal ingredients (such as those provided by category theory) or even non-formal ingredients, as required by pragmatic accounts of scientific theories.)

- A set of dynamically possible trajectories \(\mathcal{D}\) which we also call solutions of \(T\). These are those kinematically possible trajectories (\(\mathcal{D} \subset \mathcal{K}\)) which are selected by the theory’s laws in a particular application of the theory, given some choice of background structure and possibly taking into account some “nonformal patterns in theories” \cite{Cra02} p. 55].

In the next section, we will put these ingredients of a scientific theory into connection with the definitions introduced in Section 3.1. In doing so, we will have to distinguish between a general theory \(T\) and those theories which have been put forward (or are anticipated) by contemporary natural science. Similar to Chalmers’ use of the term ‘physical domain’ (Section 2) we will refer to the latter as physical theories. We will use the symbol \(T_P\) to indicate one of these theories and denote its dynamical variables, background structure, kinematically possible trajectories and solutions by \(d_P, b_P, K_P\) and \(D_P\). Finally, we will assume that the physical theories are formulated in terms of a state space \(P\), which is chosen such that according to the laws of \(T_P\), each \(p \in P\) determines a unique trajectory in \(D_P\). I.e., there is a one-to-one correspondence between solutions \((p_t)_{t \in I} \in D_P\) and states \(p \in P\).

We use the term model to denote a theory which is being proposed. This includes full-fledged theories which have not received the kind of empirical support usually required in science, but also “toy-models”, which do not aim for a comprehensive account of some class of phenomena, but rather serve to study some specific aspect of it or to test a general idea of how the phenomena could be modelled.

Finally, for use in the next section, we review the definition of a symmetry group:

**Definition 3.23.** A group \(\mathcal{G}\) is a symmetry group of a theory \(T\) \cite{Giu09} p. 43] if and only if the following conditions are satisfied:

(a) There is an effective\(^{28}\) action \(\mathcal{G} \times \mathcal{K} \rightarrow \mathcal{K}\) of \(\mathcal{G}\) on \(\mathcal{K}\).

(b) This action leaves the the solutions \(\mathcal{D}\) of \(T\) invariant.

If \(\phi\) is an action of \(\mathcal{G}\) on \(\mathcal{K}\) which satisfies the requirements (a) and (b), the pair \((\mathcal{G}, \phi)\) is a symmetry of \(T\).

### 3.4. Models of Consciousness

In Section 3.2, we have discussed that qualia as defined in Definition 3.8 cannot be explained in terms of contemporary natural science. This, however, is not in conflict with the application of the scientific method per se, as we aim to show in this section. What follows is inspired by the research program proposed in \cite{Cha96}, which defines the task of the scientific study of consciousness as finding “new fundamental laws (...) specifying how phenomenal (or proto-phenomenal) properties depend on physical properties” \cite{Cha96} p. 127].

When working with the definitions introduced in Section 3.1, there is no need to assume the physical domain as closed. Therefore, there is no need to restrict analysis

\(^{27}\)The last of these assumptions will only be used in Section 3.1 and could also be dropped.

\(^{28}\)An action is effective (\(\equiv\) faithful) if and only if no group element other than the identity fixes all elements of \(\mathcal{K}\).
to “supervenience laws” (cf. Section 2). A more proper framework for proposing a methodology for the scientific study of consciousness in the present context are scientific theories as introduced in Section 3.3. Let \( T_P \) denote a contemporary physical theory as specified at the end of Section 3.3.20

**Definition 3.24.** A pre-model of consciousness \( M \) is a theory as defined in Section 3.3 where:

(i) The dynamical variables are a Cartesian product of the physical state space \( P \) of \( T_P \) together with one copy of the label space \( L \) for each experiencing subject,

\[
d = L \times L \times \ldots \times L \times P.
\]

(ii) Kinematically possible trajectories \( K \) are a subset of families of dynamical variables,

\[
K \subset \left\{ \left( \ell^1_t, \ell^2_t, \ldots, \ell^k_t, p_t \right)_{t \in I} \right\},
\]

where \( \ell^i_t \in L, p_t \in P, k \) is the number of label spaces in (3.13) and \( I \) is some parameter space. 31

Due to the general form of theories introduced in Section 3.3, a model of consciousness \( M \) comes with laws \( L \) which select from all kinematically possible trajectories \( K \) a set of solutions \( D \). Each solution \( \left( \ell^1_t, \ell^2_t, \ldots, \ell^k_t, p_t \right)_{t \in I} \in D \) consists of families \( \left( \ell^i_t \right)_{i \in I} \), which describe changes of labels for every experiencing subject \( i \in \{1, \ldots, k\} \), and of a family \( \left( p_t \right)_{t \in I} \), which describes changes of the physical states. The solution thus realizes the mutual influence of qualia and physical states as described by the laws of the model \( M \).

3.4.1. **Notation.** We will generally use the shorthand

\[
(\bar{\ell}_t, p_t)_t := \left( \ell^1_t, \ell^2_t, \ldots, \ell^k_t, p_t \right)_{t \in I},
\]

where \( \bar{\ell}_t = (\ell^1_t, \ell^2_t, \ldots, \ell^k_t) \), to denote elements of \( K \). Furthermore, we denote by \( D|_P \) those trajectories in the physical state space \( P \) which are part of solutions \( D \) of \( M \),

\[
D|_P := \left\{ \left( p_t \right)_{t \in I} \left| \bar{\ell}_t, p_t \right)_{t \in I} \in D \right\}.
\]

This set is not necessarily equal to the set \( D_P \) of solutions of the contemporary physical theory \( T_P \). Whether \( D_P = D|_P \) or \( D_P \neq D|_P \) is determined by the laws \( L \) of the model \( M \), cf. Section 3.6.28 Similarly, we define

\[
K|_P := \left\{ \left( p_t \right)_{t \in I} \left| \bar{\ell}_t, p_t \right)_{t \in I} \in K \right\},
\]

\[
K|_L := \left\{ \left( \bar{\ell}_t \right)_{t \in I} \left| \bar{\ell}_t, p_t \right)_{t \in I} \in K \right\}.
\]

Since the choice of subset \( K \) in (3.14) is a technical condition prior to the application of any law \( L \), we may for simplicity assume that \( K|_P = K \). 31

---

20For example, \( T_P \) could refer to Quantum Mechanics with \( P \) being a particular Hilbert space \( \mathcal{H} \), or to a Quantum Field Theory, where \( P \) is a Fock space \( \mathcal{F} \), or to a model of a neural network, where \( P \) is the space of states of the network.

28The subset \( K \) will typically be determined by demanding families \( \left( \ell^1_t, \ell^2_t, \ldots, \ell^k_t, p_t \right)_{t \in I} \) to satisfy some mathematical properties, such as regularity, which are necessary for the laws \( L \) of \( T \) to be well-defined. To exclude pathological cases, we assume that every label \( \ell \in L \) is contained in at least one family \( \left( \ell^1_t, \ell^2_t, \ldots, \ell^k_t, p_t \right)_{t \in I} \in K \).

31This assumption will be used in Section 5.1 and can otherwise also be dropped.
### 3.4.2. Necessary Condition for Empirical Well-Definedness

In Section 3.1, we have discussed intersubjectively meaningful references to qualia. We have found that sequences of labels \((\ell_1, \ldots, \ell_n)\) are not empirically well-defined and have shown that the empirically well-defined references to qualia are precisely the equivalence classes \((3.8)\). We now repeat a similar analysis for pre-models of consciousness. We first introduce the necessary mathematical tools.

Let \(s \in \text{Aut}(L)\) be an element of the automorphism group \((3.5)\) which describes the freedom of an experiencing subject to choose labels for the qualia he/she experiences. Given a solution \((\bar{\ell}_t, p_t) \in \mathcal{D}\), we may apply \(s\) to that label space in \((3.13)\) which is associated to the \(i\)th experiencing subject. This gives another trajectory

\[
(\ell^1_t, \ldots, s(\ell^i_t), \ldots, \ell^k_t, p_t)_{t \in \mathbb{I}}
\]

where \(i \in \{1, \ldots, k\}\). The map which takes \((\bar{\ell}_t, p_t)\) to \((3.18)\) is an \(\text{Aut}(L)\)-action \(\phi_i\) on \(\mathcal{K}\), defined as

\[
\phi^i : \text{Aut}(L) \times \mathcal{K} \rightarrow \mathcal{K}
\]

\[
(s, (\ell^1_t, \ldots, \ell^i_t, \ldots, \ell^k_t, p_t)) \mapsto (\ell^1_t, \ldots, s(\ell^i_t), \ldots, \ell^k_t, p_t)
\]

where the subscript \(i\) indicates on which label space \(\text{Aut}(L)\) acts. We may take into account the freedom of every experiencing subject to relabel his/her qualia by considering an action \(\phi\) of

\[
\text{Aut}(L)^k := \text{Aut}(L) \times \ldots \times \text{Aut}(L)
\]

on \(\mathcal{K}\), defined as

\[
\phi : \text{Aut}(L)^k \times \mathcal{K} \rightarrow \mathcal{K}
\]

\[
(s_1, \ldots, s_k, (\ell^1_t, \ldots, \ell^i_t, \ldots, \ell^k_t, p_t)) \mapsto (s_1(\ell^1_t), \ldots, s_k(\ell^k_t), p_t)
\]

This action corresponds to the transformations we have considered in Section 3.1. However, in the context of models of consciousness, this is not the most general form of relabelling. The most general form is

\[
\sigma : \text{Aut}(L)^k \times \mathcal{K} \rightarrow \mathcal{K}
\]

\[
(\bar{s}, (\ell^1_t, \ldots, \ell^k_t, p_t)) \mapsto (s_1(\ell^1_t), \ldots, s_k(\ell^k_t), p_t')
\]

where \(p_t'\) is given by an action \(\bar{\sigma}\) of \(\text{Aut}(L)^k\) on \(\mathcal{K}|_P\),

\[
\bar{\sigma} : \text{Aut}(L)^k \times \mathcal{K}|_P \rightarrow \mathcal{K}|_P
\]

\[
(\bar{s}, (p_t)) \mapsto (\bar{p}_t')
\]

This action \(\sigma\) reduces to the action \((3.21)\) if \(\bar{\sigma}\) is trivial. If \(\bar{\sigma}\) is non-trivial, \(\sigma\) specifies that the physical states are relabelled along with the qualia. We will see below (cf. Section 3.5) for details) that the possibility of a non-trivial \(\bar{\sigma}\) is what allows us to go beyond the standard methodology explained in Section 3.1.

**Notation:** Similarly to \((3.15)\), we will use the shorthand \(\bar{s} := (s_1, \ldots, s_k)\) and also combine the two as \(\bar{s}(\ell_t) := (s_1(\ell^1_t), \ldots, s_k(\ell^k_t))\). As usual, we denote \(\sigma(\bar{s}, (\ell_t, p_t))\) as \(\sigma_{\bar{s}}((\ell_t, p_t))\). Furthermore, we use \(k := (\bar{\ell}_t, p_t) \in \mathcal{K}\).

**Remark 3.25.** We remark that each action \(\sigma\) of the form \((3.22)\) has two different meanings: On the one hand, they describe a relabelling of the trajectory \(k\). I.e., \(\sigma_{\bar{s}}(k)\) describes the same situation as \(k\) but with respect to a different choice of labelling. This is the meaning we have considered in Section 3.1. It is analogous to a change of
reference frame in physics. On the other hand, \( k' := \sigma s(k) \) is simply another trajectory in \( K \), which for \( \bar{s} \neq \text{id} \in \text{Aut}(L)^k \) describes a scenario which is genuinely different to that of \( k \). Whereas according to \( k \), at time \( t \) experiencing subject \( i \) experiences the quale he/she has labelled as \( \ell_i^t \) and physical state \( p_t \) pertains, according to \( k' \) the experiencing subject experiences a quale he/she has labelled \( s_i(\ell_i^t) \) and physical state \( p'_t \) pertains. This is reminiscent of the distinction between active and passive transformations in physics. Using this terminology we have:

1. **Passive meaning of \( \sigma \):** \( k \) and \( \sigma s(k) \) are the same trajectory expressed in different labelling.

2. **Active meaning of \( \sigma \):** \( k \) and \( \sigma s(k) \) are different trajectories expressed in the same labelling.

The fact that active and passive transformation have an identical mathematical form is related to the fact that qualia by definition cannot be referenced intersubjectively.

Next, we use the fact that \( k \) and \( \sigma s(k) \) describe the same trajectory with respect to two different choices of labels. Since a different choice of labels must not make a difference, it follows that if \( k \) is a solution of \( M \), \( \sigma s(k) \) needs to be a solution as well, for any choice of \( \bar{s} \in \text{Aut}(L)^k \). This leads us to the following definition:

**Definition 3.26.** A necessary condition for a pre-model of consciousness \( M \) to be empirically well-defined is that there is an \( \text{Aut}(L)^k \) action (3.22) on \( K \) which maps solutions to solutions, i.e. which satisfies

\[
\sigma s(D) = D
\]  

(3.24)

for all \( \bar{s} \in \text{Aut}(L)^k \).

Using Definition 3.24, this yields the following lemma.

**Lemma 3.27.** A necessary condition for a pre-model of consciousness \( M \) to be empirically well-defined is that \( \text{Aut}(L)^k \) is a symmetry group of \( M \) whose action is of the form (3.22).

This lemma establishes a connection between the laws \( L \) of the model, which determine the set of solutions, and the group \( \text{Aut}(L)^k \).

**Proof.** According to Definition 3.24 \( \text{Aut}(L)^k \) is a symmetry of the model \( M \) iff (3.22) is effective and leaves \( D \) invariant. Invariance holds by Definition 3.26. Effectivity holds because for large enough \( K \) (cf. Footnote 30) every action of the form (3.22) is effective: For any \( \bar{s} \in \text{Aut}(L)^k \) with \( \bar{s} \neq \text{id} \), there exists an \( \ell_1^t \in L \) such that \( s_i(\ell_1^t) \neq \ell_1^t \) as well as a trajectory \( k \in K \) which contains this label, so that \( \sigma s(k) \neq k \).

3.4.3. **Sufficient Condition for Empirical Well-Definedness.** In Section 3.4.2 we have found that the existence of an action (3.22) which constitutes a symmetry as defined in Definition 3.24 is a necessary condition for the solutions \( D \) of the pre-model to be empirically well-defined. We therefore include this requirement in the definition of a model of consciousness:

**Definition 3.28.** A **model of consciousness** is a pre-model of consciousness \( M \) (Definition 3.24) which carries an \( \text{Aut}(L)^k \) symmetry of the form (3.22).
Note that this is a requirement with respect to the laws $L$ of a pre-model of consciousness $M$: The laws need to be such that there is an action $\sigma$ which turns \[ p_t \] into a symmetry of $M$.

In Section 3.5 we will show that this framework indeed allows to go beyond the limitations of the standard approach explained in Section 3.1. To furthermore make the point that this is a sufficient mathematical framework for the scientific study of consciousness (i.e. sufficient to study qualia proper, cf. Definition 3.17), we show in the following example that a typical class of ideas put forward in the context of neuroscience can indeed be formalized in this framework: The idea that qualia are determined by physical states.

**Example 3.29.** In this example, we consider the hypothesis that qualia are determined by physical states, e.g. by neural activity in the brain. The naive formalization of this idea would be to consider a function (in the mathematical sense) which specifies which quale an physical states, e.g. by neural activity in the brain. The naive formalization of this idea is as in (3.22), which does not achieve the task of giving a definition of $\sigma$ as in (3.23), as we now explain.

In order to properly formalize this idea, we proceed as follows.

For simplicity, we consider the case of one experiencing subject ($k = 1$). We assume that a particular labelling has been fixed by the experiencing subject and assume that with respect to this labelling a function

$$ f : P \to L \quad p \mapsto f(p) $$

is given, where $P$ is the state space of a physical theory $T_P$ as above and $L$ denotes the label space. This function could e.g. be the result of experiments which include the experiencing subject in question. The state space $P$ could, e.g., refer to neural activity.

Based on this function $f$, we can define a pre-model of consciousness $M$. To this end, we set $d = L \times P$, choose $K$ as the right hand side of (3.14) and define the solutions of the model in terms of the solutions $\{(p_t)_t \in \mathbb{D}_P\}$ of the physical theory $T_P$ as

$$ \mathbb{D} = \{(f(p_t), p_t) \mid (p_t)_t \in \mathbb{D}_P\}. $$

The solutions of this model are thus given by the solutions of the physical theory (e.g. brain dynamics) equipped with qualia as specified by $f$.

As it stands, this model is not invariant with respect to relabelling. E.g., if the choice of labels is being changed according to some $s \in \text{Aut}(L)$, the solution $(f(p_t), p_t)_t$ is being mapped to the solution $(s(f(p_t)), p_t)_t$ which in general will not be an element of $\mathbb{D}$ as defined in (3.26). Thus the theory is not empirically well-defined.

In order to establish empirical well-definedness, there are two choices: First, one could demand that $s(f(p)) = f(p)$ for all $s \in \text{Aut}(L)$ and all $p \in P$. This amounts to considering a function $f : P \to L \backslash \sim$, where $L \backslash \sim$ is as in (3.8), which does not achieve the task of Definition 3.17. The alternative is to specify an action $\tilde{\sigma}$ as in (3.22), as we now explain.

The action $\tilde{\sigma}$ describes how the physical state changes along with a change of qualia (active interpretation, cf. Remark 3.25). We observe that a definition of $\tilde{\sigma}$ as

$$ \tilde{\sigma}_s((p_t)_t) := (p'_t)_t \quad \text{with} \quad p'_t := f^{-1}(s(f(p_t))) $$

where $f^{-1}(\ell)$ denotes any element of the pre-image $f^{-1}(\ell)$ of $\ell$, yields for (3.22)

$$ \sigma_s((f(p_t), p_t)_t) = (s(f(p_t)), p'_t)_t = (f(p'_t), p'_t)_t. $$

This includes the determination in terms of properties of the physical states or in terms of functional roles exhibited by the physical states.
where we have used \( f(p_t') = f \circ f^{-1}(s(f(p_t))) = s(f(p_t)) \). Thus if \((p_t')_t\) is a solution of \(T_P\), the action (3.22) with \(\tilde{\sigma}\) as defined in (3.27) is a symmetry of \(M\), so that \(M\) is a model of consciousness, i.e. empirically well-defined.

Thus the idea that physical states determine qualia can indeed be formalized, even though qualia are defined to be non-communicable aspects of experience. The limitation of this approach is that the function \(f\), being defined with respect to a particular choice of labelling of the experiencing subject, cannot be interpreted as specifying the quale which the experiencing subject experiences along with a particular physical state \(p\). Nevertheless, the formalism allows to treat the case that a quale, whichever one it is among the qualia in the equivalence class \([f(q)]\), is determined by the physical state \(p\). Here, the equivalence class in question is the one defined in (3.7). A further analysis of the difference to a direct description will be given in Section 3.5.

We close this section by specifying the empirically well-defined part of the trajectories of a model of consciousness \(M\). This specification is analogous to the specification of empirically well-defined sequences in (3.6).

As in Section 3.1, we define two trajectories \((\bar{\ell}_t, p_t)_t\) and \((\bar{\ell}'_t, p'_t)_t \in K\) to be equivalent if one can be obtained from the other by relabelling the qualia of the experiencing subjects. In contrast to Section 3.1, relabelling is defined in terms of the action (3.22), which for non-trivial \(\tilde{\sigma}\) includes a relabelling of the physical states. We denote this equivalence by \(\sim_\sigma\),

\[
(\bar{\ell}_t, p_t)_t \sim_\sigma (\bar{\ell}'_t, p'_t)_t \quad \text{if and only if there is an } \bar{s} \in \text{Aut}(L)^k \quad \text{such that } (\bar{\ell}'_t, p'_t)_t = \sigma_{\bar{s}}((\bar{\ell}_t, p_t)_t). \tag{3.28}
\]

Note that \(\tilde{\sigma}\), and hence \(\sigma\), depends on the laws of the model \(M\) under consideration. The empirically well-defined part of the trajectories is given by the quotient set

\[
K/\sim_\sigma, \tag{3.29}
\]

i.e. by the the space of equivalence classes of \(\sim_\sigma\). This space describes the distinctions which remain once all trajectories are identified which can be mapped to each other by relabelling the qualia of the experiencing subjects.

### 3.5. Comparison with Direct Description

Before going on, we compare models of consciousness as defined above to what may be called a ‘direct description’ of qualia: A description simply in terms of labels, without invoking the particularities of Definitions 3.24 and 3.28.

A direct description of the qualia of \(k\) experiencing subject is simply a family

\[
(\ell^1_t, \ldots, \ell^k_t)_{t \in \mathcal{I}}, \tag{3.30}
\]

where \(\mathcal{I}\) is some parameter space as above. As explained in Section 3.1, (3.30) has a large ambiguity, and the corresponding well-defined statement is given by (3.6), which in the present case reads

\[
(\bar{\ell}_t)_t \sim (\bar{\ell}'_t)_t \quad \text{if and only if there is an } \bar{s} \in \text{Aut}(L)^k \quad \text{such that } \ell^i_t' = s_i(\ell^i_t) \text{ for all } t \in \mathcal{I} \tag{3.31}
\]

and

\[
[(\bar{\ell}_t)_t] := \{(\bar{\ell}'_t)_t | (\bar{\ell}'_t)_t \sim (\bar{\ell}_t)_t\}. \tag{3.32}
\]

In order to compare this with a model of consciousness, we assume that one wishes to relate (e.g. statistically analyse) the direct description of qualia with some properties
MODELS OF CONSCIOUSNESS

of a physical system, e.g. neural activation patterns. We assume that these properties are determined by states of the physical system and denote the state space as above by \( P \). Thus the data under consideration is of the form

\[
(\ell^1_t, \ldots, \ell^k_t, p_t)_{t \in \mathcal{I}}.
\]  

(3.33)

It could result, e.g., from verbal reports of labels and simultaneous fMRI scans or EEG recordings.

In (3.29), we have found that the empirically well-defined trajectories of a model of consciousness are given by the quotient set

\[
\mathcal{K} / \sim_\sigma.
\]  

(3.34)

The following lemma gives the corresponding result for the direct description.

**Lemma 3.30.** The empirically well-defined trajectories of a direct description of qualia are given by the quotient set

\[
\mathcal{K} / \sim_\phi,
\]  

(3.35)

where \( \phi \) is the action (3.21) and where \( \sim_\phi \) is defined as in (3.28).

The difference between both sets is that in (3.34) one takes the quotient with respect to an action that generally acts non-trivially on the physical trajectories \( (p_t)_t \), whereas in (3.35) one takes the quotient with respect to an action that acts trivially on the latter. We defer further discussion to after the proof.

**Proof of Lemma 3.30.** In terms of the notation (3.33), the equivalence (3.31) is given by

\[
(\bar{\ell}^1_t, p_t)_{t} \sim (\bar{\ell}^2_t, p'_t)_{t}
\]  

if and only if there is an \( \bar{s} \in \text{Aut}(L)^k \) such that

\[
(\bar{\ell}^1_t, p_t)_{t} = \phi_{\bar{s}}((\bar{\ell}^2_t, p'_t)_{t}).
\]  

(3.36)

According to (3.28), this is precisely the equivalence \( \sim_\phi \). Hence the empirically well-defined trajectories (3.32) are elements of the quotient set (3.35).

The quotient space (3.35) gives the empirically well-defined trajectories that can be constructed from the standard approach used in contemporary scientific studies of consciousness. The quotient space (3.34), on the other hand, gives the empirically well-defined trajectories of a model of consciousness. The difference between both is determined by the action \( \bar{\sigma} \) defined in (3.23), which describes how the physical states transform if one relabels the qualia of an experiencing subject (passive meaning, cf. Remark 3.25), but also how the physical states change if the qualia of an experiencing subject change (active meaning in Remark 3.25). It is precisely the possibility of a non-trivial \( \bar{\sigma} \) which allows the laws postulated by a model of consciousness to address individual qualia, i.e. “facts that embody a particular point of view” [Nag74, p. 441], to a certain extend. The extend to which this is possible is limited by the requirement that (3.22) constitutes a symmetry of the Model \( M \).

Mathematically, this is reflected in the fact that elements of (3.35) are of the form

\[
([\ell^1_t]), \ldots, ([\ell^k_t]), (p_t)_{t},
\]  

(3.37)

where each equivalence class \([\ell^i_t]\) is as in (3.31) and (3.32). This is precisely what can be studied by usual means if there are \( k \) experiencing subjects (cf. quotient space (3.3)). The elements of (3.34), on the other hand, are not of this form: A non-trivial \( \bar{\sigma} \) results in equivalence classes in which the physical states are *mixed* with the
labels of the various experiencing subjects. The empirically well-defined trajectories cannot be separated as in \((3.37)\).

A simple way to put all of this is as follows. If one chooses a direct description of qualia and investigates the relation to the physical domain, one first has to consider equivalence classes \((3.32)\) (for only they are empirically well-defined) and subsequently propose or analyse the relation to the physical domain. Using a model of consciousness, one may exchange this order: One may first postulate a relation of qualia and the physical domain, and subsequently remove the arbitrariness of choosing labels by considering equivalence classes \((3.28)\). The requirement of there being a symmetry \(\sigma\) simply makes sure that this relation (the law \(L\) of a model of consciousness) is chosen in such a way that the second step – obtaining an empirically well-defined theory – is possible at all. A non-trivial \(\tilde{\sigma}\) implies that after the second step, one does not end up with what one could have obtained using the first procedure right away. This small detail can have large empirical consequences, as we show e.g. in Section \(\text{[5]}\).

Note that since \(\tilde{\sigma}\) may be chosen as trivial, direct descriptions constitute a special type of model of consciousness and the latter are in fact a generalization of the former. What we have shown above is that direct descriptions, nevertheless, are not suitable for the task specified in Definition \(\text{[3.17]}\).

This concludes the definition and analysis of models of consciousness. In next section, we introduce one last important concept: The closure of the physical. Subsequently, we close this part of the paper by giving several examples in Section \(\text{[4]}\).

3.6. **Closure of Physical.** The closure of the physical domain, roughly the idea that “physical laws already form a closed system” \([\text{Cha10, p. 17]}\), is an important conception which underlies many philosophical investigations of consciousness. In Section \(\text{[2]}\) we have explained that a specific version of it is crucial for Chalmers’ grounding of the scientific study of consciousness, which limits the applicability of this grounding (cf. Appendix \(\text{[A]}\)). This is different for the phenomenological grounding.

Since neither the basic definitions (Section \(\text{[3.1]}\), nor models of consciousness (Definition \(\text{[3.28]}\) refer to the closure of the physical in any way, this is an independent assumption which one may or may not make when working in the phenomenological grounding. This assumption in fact takes a very simple mathematical form:

**Definition 3.31.** A model of consciousness \(M\) describes the *physical as closed* if and only if
\[
\mathcal{D} \cap P = \mathcal{D}_{\mathcal{P}},
\]
where \(\mathcal{D} \cap P\) is defined in \((3.16)\) and where \(\mathcal{D}_{\mathcal{P}}\) has been introduced in Section \(\text{[3.3]}\) to denote the solutions of the physical theory \(T\) which underlies the model \(M\).

In words, this definition says that a model of consciousness \(M\) describes the physical as closed if and only if the physical trajectories which are determined by the laws of \(M\) (as part of the solutions \(\mathcal{D}\)) are, as a set, equal to the solutions of the physical theory \(T\) which \(M\) is based on. Whether or not \((3.38)\) is satisfied depends on the laws \(L\) postulated by a particular model.

The closure of the physical, just like its opposite, are metaphysical assumptions. To date, there is no valid argument which shows that either of them cannot hold in nature (cf. Section \(\text{[A.1]}\)). However, a transcendental argument, which we explain in detail in Remark \(\text{[A.1]}\) implies that one of these assumption is not compatible with a scientific approach to consciousness as defined by Chalmers’ grounding or the phenomenological
grounding. Surprisingly, this is the assumption of the closure of the physical. In a nutshell, this is so because this assumption is very strong, strong enough to say something about communication channels or storage devices which are involved in any scientific activity, in particular in experiments (Section A.2).

If Remark A.1 holds true, this implies, among other things, that one can either assume the closure of the physical and admit that the relation of qualia to the physical domain is not amenable to an empirically meaningful scientific activity, or restrict attention to models of consciousness which do not describe the physical as closed, i.e. to models where some qualia “influence” the physical domain (in the sense of Definition 3.31).

Remark 3.32. At this point it is essential to discuss again the class $C$ of experiencing subjects which has been introduced in Section 3.1, because this class determines which aspects of experience satisfy the definition of qualia, and which not. If this class is chosen as very large, e.g. comprising humans as well as bats, many aspects of experience of humans will be experienced as non-communicable, simply because of a general lack of means of communication. In other words: The set of aspects of experience which satisfy Definition 3.8 will largely exceed those aspects typically referred to in the context of the mind-body problem (e.g. aspects of the ‘what it is like to be’ type, cf. Example 3.10). Whereas everything that has been said in previous sections still holds true for this case, this point becomes important when having to make metaphysical choices, such as the one related to the closure of the physical.

In order to avoid this problem, one could demand that the class $C$ is chosen as small as possible (comprising at least two experiencing subjects, and more if necessary, e.g. due to statistical considerations). However, eventually, further considerations are necessary to decide which qualia should influence the physical domain, and which not.

4. Examples

In this section, we review some theories and models about consciousness or the mind-matter relation which have been proposed in the literature. Our goal is two-fold: On the one hand, we aim to illustrate the large amount of thought and effort which has gone into the construction and study of formal approaches to the mind-matter relation. On the other hand, we aim to evaluate which of the existing models are compatible with the phenomenological grounding, and which not. Needless to say, this list of examples is not complete.

4.1. Integrated Information Theory. Our first example is Integrated Information Theory (IIT) which has been proposed by Giulio Tononi in 2004 [Ton08] and has since been developed considerably. The current version of that theory [OAT14] consists of an algorithm whose input is a model of a physical system (together with a state of that system and including its dynamical laws) and whose output are formal quantities which give answers to the following three questions: 1. Which parts of the system are conscious? 2. What are they conscious of? 3. How conscious are they?

To answer the first question, the theory’s algorithm identifies some (mutually disjoint) subsystems of the system. To answer the second question, for each such subsystem $S$, the algorithm specifies what is called a ‘maximally irreducible conceptual structure’ (MICS). This is a mathematical object of the following kind: Let $\mathcal{P}_S$ be
the space of probability distributions (or probability measures) over the states of the subsystem $S$. A ‘concept’ is an element of the space $^{34}$

$$\mathcal{P}_{CS} := \mathcal{P}_S \times \mathcal{P}_S \times \mathbb{R}_0^+.$$  

The maximally irreducible conceptual structure is an $n$-tuple of concepts, where $n$ is determined dynamically by the theory and may vary from subsystem to subsystem. I.e., it is an element of the “qualia space” [OAT14, graphical illustration in Figure 15]

$$L_S := \mathcal{P}_{CS}^{\times n(S)}.$$  

Finally, in order to answer the third question, the algorithm specifies the integrated conceptual information $\Phi_{\text{max}}(S) \in \mathbb{R}_0^+$. In summary:

“[T]he central identity [of IIT] is the following: The maximally irreducible conceptual structure (MICS) generated by a [subsystem $S$] is identical to its experience. The constellation of concepts of the MICS completely specifies the quality of the experience (...). Its irreducibility $\Phi_{\text{max}}$ specifies its quantity.” [OAT14, p. 3].

The main papers of the theory remain somewhat silent about what exactly they take the terms “consciousness” or “quality and quantity of an experience” [OAT14] to mean. However, the theory is often taken as an example of Chalmers’ grounding of the scientific study of consciousness, so that (4.1) is taken to be the space of qualia as defined in (D1), and the algorithm specified by IIT is taken to constitute a ‘psychophysical law’. To evaluate whether or not the former identification, i.e. the central identity of IIT, is plausible if one defines qualia according to Chalmers’ grounding goes beyond the scope of this example. What concerns us here is to evaluate in how far the theory fits with the phenomenological grounding put forward above.

The first observation is that ineffable aspects of conscious experience seem to have played at least a small role in the early development of the theory. E.g., in [Ton08, p. 229], Tononi notes that “[t]he notions just sketched aim at providing a framework for translating the seemingly ineffable qualitative properties of phenomenology into the language of mathematics” (our emphasis). As we have explained in detail in Section 3.1, ineffable aspects of experience cannot be put in a one-to-one correspondence with mathematical objects, simply because two or more experiencing subjects have no means to ensure that they have associated the same ineffable aspect of experience with the same mathematical object. This was the reason for us to introduce label spaces $L$ in Section 3.1 and what lead to the requirement of there being a symmetry that describes relabelling. Following this path, we might take IIT’s “qualia space” $L_S$ to constitute the label space of qualia as defined in Definition 3.8.

This brings us to the question of how communicable relations between qualia so defined, such as the ones put forward in Phenomenological Fact 3.5, are related to the mathematical structure of the space $L_S$. First, we note that the space $L_S$ can be equipped with a metric: For any metric $d$ on $\mathcal{P}_S$ and using the usual metric on $\mathbb{R}_0^+$, summation allows to define both a metric on $\mathcal{P}_{CS}$ and $L_S$. This metric can be taken to express the similarity relation in Phenomenological Fact 3.5. And indeed, this may again have been a guiding idea in the development of the theory, “experiences are similar if their shape is similar” [Ton08, p. 228]. Next, what has been called

---

$^{34}$In the terminology of OAT14, a concept consists of the maximally-irreducible cause and effect repertoires of a mechanism $M$ of $S$ together with its integrated information $\varphi(M)$, provided that the latter is non-zero [MMA+18, Supplementary S1, p. 176].
the “intensity” of an experience in Example 3.14 corresponds to the “quantity” of experience according to IIT. The corresponding mathematical structure is the $\mathbb{R}^+_0$ in which $\Phi_{\text{max}}(S)$ takes values.

IIT arguably encodes another communicable relation between qualia: Their composition in experience. This is usually formulated as an axiom of IIT which states that “[c]onsciousness is compositional (structured): each experience consists of multiple aspects in various combinations” [OAT14, p. 2]. The composition of the experience of a subsystem $S$ in terms of more elementary experiences of the same subsystem is modelled by the Cartesian product structure of (4.1) in terms of the concept spaces $\mathcal{P}_{CS}$. One may interpret the $\mathbb{R}^+_0$ that constitutes the last factor of $\mathcal{P}_{CS}$ as the intensity of the more elementary experiences.

In summary, the basic definitions of IIT seem to fit quite well with the basic definition of the phenomenological grounding and IIT can be taken to constitute a pre-model of consciousness as defined in Definition 3.24. However, in order to take into account the non-communicability of the corresponding aspects of experience, the symmetry (3.22) has to be implemented. This can be done by simply swapping the states of the physical system that give rise to particular labels $\ell \in L_S$: If $\ell_1, ..., \ell_n \in L_S$ are the labels of the conscious subsystems of a system in state $p_1$, and $\ell'_1, ..., \ell'_n \in L_S$ are the labels of conscious subsystems of the system in state $p_2$, for any $\bar{s} \in \text{Aut}(L_S)$ which maps the former to the latter, we define the action (3.23) as $\tilde{\sigma}_{\bar{s}}(p_1) = p_2$. (For details, cf. Example 3.29.) Equipped with this symmetry, IIT constitutes a model of consciousness as defined in Definition 3.28.

We conclude this example with some conceptual remarks. First of all, we note that the phenomenological grounding approaches model-building differently than [OAT14]. Whereas in the latter, phenomenological facts are used to justify the definition of the algorithm, i.e. the dynamical equations of IIT, the phenomenological grounding uses phenomenological facts to model the mathematical space associated with qualia. We have seen above that an earlier version of IIT put forward in [Ton08] is more aligned with this perspective. From the perspective of the phenomenological grounding, the main task would be to motivate the mathematical structure of (4.1) in more detail. E.g., one could ask why the elementary qualia are labelled by elements of the space $\mathcal{P}_S \times \mathcal{P}_S$ and not by a simpler metric space? Does the structure of the former have any phenomenological interpretation? Questions of this sort may have large consequences for the further development of the theory because the algorithm of IIT in its current form makes essential use of elements of $\mathcal{P}_S \times \mathcal{P}_S$.

Second, we remark that IIT describes the physical as closed: The dynamical evolution of the physical domain is not changed in any way by the theory. Thus, if one interprets the theory in Chalmers’ or the phenomenological grounding (or similar ones, in fact) the question arises of how the theory’s mathematical postulates – first and foremost the algorithm it specifies – can be evaluated experimentally at all, for it seems that all results one can hope to obtain from neuroscientific experiments that scan the brain (EEG, fMRI, etc) are physical and therefore determined by the physical domain alone. Put in simple terms, one may ask what one actually learns when collecting experimental data that is in fact completely determined by the physical domain. This problem seems to appear in any grounding which exhibits an explanatory gap. One could try to avoid the problem by interpreting (4.1) in terms of aspects of experience that do not exhibit an explanatory gap. This, however, would raise the question of why a novel law (the algorithm of IIT) should determine those aspects, as compared to some form of neural processing.
argument is outlined in more detail in Appendix A.2 and related to the transcendental argument against the closure of the physical given in Remark A.1. In Section 4.1.1 we review a modification of IIT which avoids this problem.

4.1.1. Integrated Information-Induced Quantum Collapse. To overcome the problem mentioned in the last paragraph, one needs to propose a model of consciousness which does not postulate the physical as closed, i.e. for which (3.38) does not hold. A first model of this kind based on Integrated Information Theory (IIT) is given in [KR15]. It refers to an early version of IIT which only answers the third question of Section 4.1: How conscious is a physical system?

The physical theory on which this model is based is a quantum system with Hilbert space $\mathcal{H}$ and Hamiltonian $H$. The states are density matrices $\rho \in L(\mathcal{H})$. Given a density matrix $\rho$, the Quantum Integrated Information is defined as

$$\Phi(\rho) = \inf \left\{ S(\rho\|\otimes_{i=1}^{N} \text{Tr}_i \rho) \right\},$$

where the infimum is taken over all decompositions of the Hilbert space $\mathcal{H}$, i.e. over all isomorphisms between $\mathcal{H}$ and a Hilbert space of the form $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_N$, where $\text{Tr}_i \rho$ denotes the reduced density matrix on the Hilbert space $\mathcal{H}_i$ (i.e. $\text{Tr}_i$ is a trace over all $\mathcal{H}_j$ with $j \neq i$) and where finally $S(\rho\|\rho')$ denotes the quantum relative entropy defined as

$$S(\rho\|\rho') = \text{Tr} \rho \log \rho - \text{Tr} \rho \log \rho'.$$

Here, $\text{Tr}$ denotes the trace over the whole Hilbert space $\mathcal{H}$. According to this model, (4.2) specifies how conscious the physical system is in the state $\rho$.

In order to specify how consciousness in turn influences the physical domain, the model modifies the time-evolution of the physical system. Whereas in quantum theory, the time-evolution of a closed system with Hamiltonian $H$ is determined by

$$\frac{\partial \rho}{\partial t} = -\frac{i}{\hbar} [H, \rho],$$

this model proposes the evolution equation

$$\frac{\partial \rho}{\partial t} = -\frac{i}{\hbar} [H, \rho] + \sum_{n,m=1}^{N^2-1} h_{n,m}(\Phi(\rho)) \cdot (L_n \rho L_m^\dagger - \frac{1}{2} \rho L_m^\dagger L_n + \frac{1}{2} L_m^\dagger L_n \rho),$$

where $h_{n,m}(\Phi(\rho))$ are continuous functions of $\Phi(\rho)$ which vanish if $\Phi(\rho) = 0$ and where $L_k$ are operators on $H$. This is a Lindblad evolution equation which describes, among other things, models of spontaneous wave function collapse. By choosing the functions $h_{n,m}$ suitably small, one can make sure that the model is compatible with physical experiments to date. The model is furthermore experimentally accessible in that it predicts a collapse rate which is dependent on $\Phi(\rho)$, rather than mass or the number of particles alone, as is the case in other spontaneous collapse models (cf. [KR15, Sec. 5]).

4.2. Global Neuronal Workspace Theory. The Global Neuronal Workspace model (GNW) is, next to Integrated Information Theory, the other model largely favoured by neuroscientists. In contrast to the latter, however, it is usually stated directly in terms of brain physiology [DCNT11, DKC98]. Even though this is sufficient to make some specific predictions [DCN11], a more formal model would be desirable, not least to make a detailed comparison with IIT possible.
In what follows, we outline how a formal model could be constructed which takes as input any physical system (in a certain class of systems) and determines what the system is conscious of. To this end, we apply concepts from dynamical systems and nonbinary information processing whose connection with consciousness has recently been suggested in [Gri18]. While this attempt is very preliminary, the hope is that a genuine formal model can be developed along these lines in future work. A different goal is pursued in [Wal05].

Let $S$ be a physical system. We assume that it consists of a set $N_v$ of components (‘vertices’, representing neurons in a neuronal network), each of which is in a particular state $u_i(t)$. Here, $t \in T$ denotes time and $i \in N_v$ denotes the component in question. Furthermore, we assume that it consists of a set $N_e$ of directed edges (representing axons, dendrites, synapses, etc. in a neuronal network), each of which may be in a state $w_l(t)$, $l \in N_e$ (e.g. representing the synaptic strength in a neuronal network). As usual, we define the parents $Pa_i$ of the $i$th component to be those components from which a directed edge leads to $i$, and assume $i \in Pa_i$. Finally, the dynamics of the system are specified component-wise by a set of ‘update-rules’ $(f_i)_{i \in N_v}$, where $f_i$ determines how the state $u_i(t)$ of the $i$th component depends on the states of its parents and the states of the edges coming from its parents at previous times.

This system has conscious representations, according to the GNW model, if two necessary conditions are satisfied. The first of these is that the system has “two main computational spaces, each characterized by a distinct pattern of connectivity” [DCN11 p.56]. The first computational space is a “processing network, composed of a set of parallel, distributed and functionally specialized processors or modular subsystems subsumed by topologically distinct (...) domains with highly specific local or medium-range connections” [ibid.]. The second computational space is a “global neuronal workspace, consisting of a distributed set of (...) neurons characterized by their ability to receive from and send back to homologous neurons in other (...) areas horizontal projections through long-range excitatory axons” [DCN11 p.56].

In order to construct a formal model of consciousness based on this hypothesis, a definition has to be given which specifies which structure in a physical system counts as a computational space of each kind, and which not; i.e. a definition of the necessary “patterns of connectivity” in terms of the mathematical structure of a physical system. In order to propose such a definition, we combine the ideas of the GNW model with some of the ideas put forward in [Gri18], using in particular the similarity between the processing network described by GNW and the “global directed network consisting of a large sparsely connected array of much smaller, irreducible subgraphs (ISGs), representing directed neuron-to-neuron connections” put forward in connection with consciousness in [Gri18]. Here, an ISG is defined as follows.

[D1] A set $N_{ISG} \subset N_v$ of components of the physical system constitutes an irreducible subgraph (ISG) if there is a directed edge between any ordered pair of components in this set [Gri18 p.25].

The similarity to the processing network of the GNW model comes about due to the major observations in [Gri18] that each ISG “acts as an analog filter, a dynamical decision-maker (preferring one or another resonant mode), an amplifier, and a router” [Gri18 p.27]. In order to specify a necessary pattern for the global neuronal workspace, we simply require that this is a network with a directed edge going into and coming out of each ISG, noting that a further requirement on this network will be
added below. In summary, a proposal for the first necessary condition for the system $S$ to be conscious may be put as follows.

[N1] The system $S$ needs to contain two disjoint subsets $N_p, N_g \subset N_v$ of components: First, a set $N_p$ of components whose induced subnetwork is a network of ISGs, were the inter-ISG-connections are feed-forward only. Second, a set $N_g$ of components with directed edges going from this set into all ISGs, and directed edges going to this set from all ISGs.

Clearly, this definition is preliminary and will have to be improved substantially to allow for a full-fledged model to be defined.

In [Gri18], general properties of the state of ISGs are explained, which have been found in previous work. In particular, if the system satisfies a few conditions, including the feed-forward property mentioned in definition [N1], the ISGs will typically carry out “successive pattern recognition tasks exploiting both remembered contextual information and prior expectations from past events (...) as well as the assumption of the structures (elements) that are identified at [a] previous level” [Gri18, p. 30]. The result of this task is recorded by a dynamical attractor on the ISG’s components, which we denote by $m_k(t)$, where $k$ indexes the ISGs in the system. These dynamical attractors represent “perceived sources/objects, (...) events, (...) narratives, (...) scenarios” [Gri18, Fig. 2]. Combining these results with the idea that “[t]he entire workspace is globally interconnected in such a way that only one such conscious representation can be active at any given time” [DCN11, p. 58], we arrive at a proposal for the second necessary condition for the system to be conscious:

[N2] The induced subnetwork of $N_g$ needs to be such that at any time $t$, its state ‘represents’ only one of the ISGs’ dynamical attractors $m_k(t)$.

Here, one could, e.g., define the term “represent” to mean that at any time $t$ the state of the network is (essentially) equal to one of the states $m_k(t)$, but other more realistic choices might be possible.

If both necessary conditions [N1] and [N2] are satisfied at a particular time $t$, the GNW model claims that the system $S$ is conscious of the “perceived object, event, narrative or scenario” $m_k(t)$ represented in the global workspace network $N_g$. Due to the directed edges from $N_g$ to the ISGs, the state of the ISG $k$ may be made “directly available in its original format to all other workspace processes” [DN01, p. 15].

Clearly, this outline leaves open various questions. Most notably, the question of how modes $m_k(t)$ of ISGs may relate to experience. Whereas IIT’s qualia space [14] has some structure which relates to phenomenology, it is highly questionable whether this can be asserted of the states of ISGs, which behave generically like a small number of “monotonically increasing phase variables” [Gri18, p. 25]. This very question arises also, albeit in a more indistinct form, if GNW is formulated in terms of neuronal architecture: How does a “piece of information selected for its salience or relevance to current goals” [DCN11, p. 56], which is really just a state of some subset of the brain’s neurons, relate to experience? A proposal to this extend is presented in Section 4.3.

Whereas from a formal modelling perspective, there is some space for further development of the GNW model, it does seem to capture essential neuroscientific evidence in a simple and very plausible hypothesis: The global workspace. This idea might ultimately be combined with ideas of IIT or other models to give an explicit account of how the state of the global workspace relates to experience as we find it.
4.3. Conscious Agent Networks. A model which is based on idealistic metaphysics is developed in [HP14]. The underlying idea is that what exists are interacting conscious agents, each of which has a fundamental capacity to perceive, decide and act, and that the interaction between these conscious agents seems to each as if there is an external outside world. For simplicity, in what follows, we explain a slight more general version of the model than presented in [HP14].

In order to explain a single conscious agent $C$, we first assume that there is a space $W$ which is external to the conscious agent. One may think of this as states of some “world” which the agent perceives, but in fact this space is constituted via interactions with other conscious agents, as explained below. Given this space $W$, a conscious agent is modelled as a five-tuple $C := (X, G, P, D, A)$, where $X$ and $G$ are spaces, and $P, D, A$ are maps interpreted as follows:

- $X$ is a space which describes possible experiences of the conscious agent. Each element $x \in X$ represents a particular experience.
- $G$ is a space which describes dispositions or intentions to act. Each element $g \in G$ corresponds to an action the agent has decided to carry out.
- $P : W \rightarrow X$ is a map which describes the agent’s “process of perception” [HP14, p. 6]. It specifies what the conscious agent experiences in response to the “world” being in a particular state $w \in W$.
- $D : X \rightarrow G$ is a map which models how the experience of the agent determines its disposition for an action, i.e. “the process of decision [in which] a conscious agent chooses what actions to take based on the conscious experiences it has.” (ibid.).
- $A : G \rightarrow W$ describes how the agent’s disposition for an action “is carried out”, i.e. how it affects the world: “In the process of action, the conscious agent interacts with the world in light of the decision it has taken, and affects the state of the world” (ibid.).

The structure of the spaces $W$ and $X$, as well as the definitions of the maps $P, G$ and $A$ are not fixed by the theory, but need to be chosen according to the application. Based on such a choice, the model specifies the dynamically possible trajectories $(x(t), g(t), w(t))_{t \in \mathbb{I}}$ as those trajectories which satisfy

$$(x(t+1), g(t+1), w(t+1)) = (Pw(t), Dg(t), Aw(t)),$$

where $t$ is chosen as a discrete time parameter, i.e. $\mathbb{I} := \mathbb{Z}$.

The central hypothesis of this theory is called “conscious realism”: That “[t]he world $W$ consists entirely of conscious agents” [HP14, p. 7]. This hypothesis is implemented via networks of conscious agents.

In order to describe a network of $k$ conscious agents, we first assume that for every conscious agent, a space $X_i$ of possible experiences and a space $G_i$ of dispositions to act is given, as well as a “decision map” $D_i$ as introduced above. The “external world” of the $i$th conscious agent is defined to be the product of the action spaces of all other conscious agents, i.e.

$$W_i := G_1 \times ... \times G_{i-1} \times G_{i+1} \times ... \times G_k.$$
This choice is motivated by the idealistic idea that what exists are only experiences and dispositions to act, and that the dispositions to act of some agents determines the experience of others. I.e., the process of perception of the $i$th conscious agent is, in case of a network of conscious agents, given by a map

$$P_i : W_i \rightarrow X_i.$$ 

This allows to define the dynamically possible trajectories of the network of conscious agents via

$$x_i(t + 1) = P_i(g_1(t), \ldots, g_k(t)) \quad \text{and} \quad g_i(t + 1) = D_i x_i(t).$$

If $P_i$ is a partial function defined only for some $G_j \in W_i$, the $i$th agent is only able to perceive the dispositions to act of the corresponding other conscious agents. Various concrete proposals for how to choose $P_i$ are discussed in [HP14, p. 7ff.].

Due to the identification of the “outside worlds” $W_i$ of each conscious agent with the dispositions to act of others, the action map $A_i : G_i \rightarrow W_i$ is not necessary to define the dynamics. In order to satisfy the definition of a conscious agent given above one may define it formally as the map which takes $g_i(t)$ to $w_i(t + 1) = (g_1(t + 1), \ldots, g_k(t + 1))$. In simple cases (e.g. involving two conscious agents [HP14]) this definition can be flushed out in terms of combinations of inverses of $D$ and $P$. In general, it may require $A_i$ to be time-dependent.

In summary, the various objects the theory assumes in a particular application determine (possibly in a probabilistic manner) the dynamics of a network of conscious agents. The goal, then, is to specify plausible assumptions which allow to deduce formally that “the perception of objects and space-time can emerge from such dynamics” [HP14, p. 1] and to “explore [the model's] theoretical implications in the normal scientific manner to see if they comport well with existing data and theories, and make predictions that are novel, interesting and testable” [HP14, p. 7].

An early example of a result of this kind is given in [HP14, p. 13ff.]. In a nutshell, it is shown that if the state spaces $X_i$ and $G_i$ are finite, the dynamics of a network of two conscious agents can be described in terms of an object which bears some similarity to a quantum-mechanical wave function of a free particle.

From the perspective of models of consciousness as defined in Section 3.4 two crucial questions arise:

a) Whether the model would like to address aspects of experience which are non-collatable.

b) Whether the theory would (eventually or in principle) like to make predictions with respect to experiments which involve (reports of) conscious agents.

An affirmative answer to the first question might be indicated by the remark that the “qualia $X$ of a conscious agent $C$ are private, in the sense that no other conscious agent $C_i$ can directly experience $X$” [HP14, p. 14]. If this is indeed the case, the mathematical structure of the spaces $X_i$ (and possibly also of $G_i$, if one holds that intentions to act are also experiences of some sort) could be defined based on a phenomenological analysis as explained in Section 3.4. This would, in particular, dismantle the objection that the “definition of conscious agents could equally well-apply to unconscious agents [so that the] theory says nothing about consciousness” [HP14, p. 14].

More importantly, if the theory also answers affirmatively to the second question, the results of Section 3.4 show that a further mathematical structure is necessary to ensure that the model is empirically well-defined (Lemma 3.27). Since conscious agent...
networks intend to link to contemporary physics proper, but do not intend to describe
the physical as closed (cf. Section 3.6), strong mathematical implications follow from
Proposition 5.10 and Corollary 5.12. These may have empirical consequences and
might offer a way to “explore [the model’s] theoretical implications” [HP14, p. 7] on
the physical domain which is independent of the particular type of network considered.

4.4. Expected Float Entropy Minimisation. One of the largest questions at present
left open by the GNW model (Section 4.2) is how the state of the global neuronal
workspace, ultimately a collection of states of individual neurons, relates to experi-
ence. Questions of this kind are addressed by the Expected Float Entropy (EFE)
model developed in [Mas16]. In short, this is a proposal for how (probability distribu-
tions of) brain states determine relations among qualia.

In what follows, we review the definition of this model. Every brain state is assumed
to consist of individual elements, each of which can be in a particular state. We denote
the set of all elements (“nodes”) by $S$ and the space of states of each node by $V$, and
assume both are a finite set. A brain state is thus a map

$$ s : S \to V. $$

(4.3)

E.g., in a neural network, $S$ is the set of neurons and $V$ is the set of possible states
of each neuron. If applied to the GNW model as outlined above, $S$ is the set $N_g$ of
nodes and $V$ denotes the corresponding space of states. In [Mas16], $s$ is called a “data
element”, but we will refer to $s$ simply as ‘state’.

Let $\Omega_{S,V}$ denote the space of all states. We assume that a probability d istribution
$p$ is given over $\Omega_{S,V}$. The probability $p(s)$ can be interpreted as the probability of
the brain being in state $s$.

A weighted relation on a set $S$ is a map $R : S \times S \to [0,1]$. Given a set of states
with corresponding probability distribution, the theory developed in [Mas16] allows to
determine two weighted relations $R$ and $U$, where $R$ is a weighted relation on the set $S$ of nodes and where $U$ is a weighted relation on the possible states $V$ of each node. We will discuss the interpretation of $R$ and $U$ at the end of this example.

The theory determines both $U$ and $R$ follows. For any state $s \in \Omega_{S,V}$, the composi-
tion $U(s(\cdot), s(\cdot))$ is a relation on $S$, which we denote as $U \circ s$. Define the float entropy $fe$ and expected float entropy $efe$ as

$$ fe(R, U, s) = \log_2 \left| \{ \tilde{s} \in \Omega_{S,V} \mid d(R, U \circ \tilde{s}) \leq d(R, U \circ s) \} \right| $$

(4.4)

$$ efe(R, U, p) = \sum_{s \in \Omega_{S,V}} p(s) fe(R, U, s) $$

(4.5)

where $s \in \Omega_{S,V}$, $d$ is a distance function on the weighted relations on $S$ and where $|A|

denotes the cardinality of a set $A$. The theory proposes “that a system (such as the
brain and its subregions) will define $U$ and $R$ (up to a certain resolution) under the
requirement that the efe is minimized.” I.e. $U$ and $R$ are defined via

$$ efe(R, U, p) = \min_{\hat{R}, \hat{U}} efe(\hat{R}, \hat{U}, p), $$

(4.6)

where the minimum is taken over all relations $\hat{R}$ on $S$ and all relations $\hat{U}$ on $V$.
(Existence or uniqueness of minimizers is not discussed in [Mas16].)

Concerning the interpretation of $R$ and $U$, the theory proposes that if “a brain state
is interpreted in the context of all these relations (...)”, the brain state acquires
meaning in the form of the relational content of the experience”. If applied to the
visual cortex, the theory aims to explain “perceived relationships between different colours, the perceived relationships between different brightnesses, and the perceived relationships between different points in a person’s field of view (giving geometry)”.

These interpretations are supported by several examples in [Mas16], where the theory is applied to pictures, so that $S$ is the set of all pixels and $V$ describes the possible colour values at each pixel, which implies that $U$ is a relation between colour values and $R$ is a relation between pixels. The support for these interpretations becomes more difficult when the theory is being applied, e.g., to the visual cortex, for in this case $U$ is a relation on the states of the neurons in the visual cortex and $R$ is a relation on the set of neurons, making it somewhat unclear why in this case a relation $U$ might give an explanation of, e.g., why “blue appears similar to turquoise but different to red”.

One can, however, simply take the theory at face value by accepting that the relata of $U$ and $R$, whichever mathematical form they take, are (describing) non-communicable aspects of experience and that $U$ and $R$ are (describing) the relations between them. Here, the non-communicability is essential for otherwise the identity of some communicable aspect of experience and elements of the set $V$ or $S$ would be questionable. In short, one may assume that the relations $R$ and $U$ correspond to the structure of the label space $L$ which describes experience.

Several interesting questions are raised by this model. First of all, we note that since the model aims to explain the relations between aspects of experience, it is fully compatible with a direct description of qualia as discussed in Section 3.5 and does not aim for a description of qualia sensu stricto. This raises the question of whether this model is an alternative to, or rather a complement of, models which do intent to describe qualia sensu stricto, such as e.g. Integrated Information Theory.

In personal communication, the author has expressed the opinion that the relations among aspects of experience might in fact be nuanced enough to allow to identify individual qualia by specification of the relations. In other words, that all orbits of the automorphism group (3.5) are trivial. Whether or not this is the case is a phenomenological question, which needs to be answered by a systematic account of the relations between qualia found in experience and is a priori to any model-building process (just like general properties of an explanandum have to be fixed prior to an explanation). However, since the EFE model actually specifies the relations between aspects of experience, one can also study which answer the model itself gives to this question. The upshot of this analysis, which is presented in the next paragraph, is that if the probability distribution $p$ is invariant with respect to a transformation (or permutation) of states, which is often the case, the model does in fact specify relations whose automorphism group has non-trivial orbits.

Consider a bijective transformation (permutation) of states $\sigma : \Omega_{S,V} \rightarrow \Omega_{S,V}$ which can be specified in terms of a bijective transformation $\sigma_S : S \rightarrow S$ of nodes and in terms of a bijective transformation $\sigma_V : V \rightarrow V$ of node-states, i.e. $\sigma(s) := \sigma_V \circ s \circ \sigma_S$. The probability distribution $p$ is invariant with respect to this transformation if $p = p \circ \sigma$, i.e. if the transformation maps states $s$ to states $\sigma(s)$ which have the same probability as the former, $p(\sigma(s)) = p(s)$. Defining the transformation of the relations $U$ and $R$ as

\[
U'(.,.) := U(\sigma_V^{-1}(.),\sigma_V^{-1}(.),) \quad \text{and} \quad R'(.,.) := R(\sigma_S(.,),\sigma_S(.,)), \quad (4.7)
\]
and using the fact that the metric $d$ is chosen as one of the $d_n$ metrics in [Mas16, p. 127], i.e. involves summation over all elements of $S \times S$, (4.4) yields that $f(e(R, U, s) = f(e(R', U', \sigma(s)))$. Using the invariance of $p$ and (4.5), this gives

$$efe(R, U, p) = efe(R', U', p).$$

This implies that for any minimizer $R, U$ of (4.4), the pair $R', U'$ is a minimizer of (4.6) as well. In other words, the theory only determines minimizers up to transformations (4.7). Assuming uniqueness of minimizers, this in turn implies that the minimizing pair $U, V$ satisfies

$$U(\cdot, \cdot) = U(\sigma_V(\cdot), \sigma_V(\cdot)) \quad \text{and} \quad R(\cdot, \cdot) = R(\sigma_S(\cdot), \sigma_S(\cdot)),$$

so that $\sigma_V$ and $\sigma_S$ are relation-preserving bijections, i.e. non-trivial elements of the automorphism group of the spaces $(V, U)$ and $(S, R)$, respectively.

Another interesting question is which part of the brain generates those relations between aspects of experience which we find in experience. This is, to a large extend, a question which could be answered by simulations of the brain’s neuronal network. If it turns out that these relations can be reproduced better by a distributed network, this model may actually be compatible, or even taken as support of, the Global Neuronal Workspace hypothesis. The underlying challenge here is, of course, to identify the weighted relations $R$ and $U$ between (states of) neurons with the manifold relations between aspects of experience. This identification may also hinge on how the probability distributions $p(s)$, which is the only data which enters the definition of $R$ and $U$, is interpreted when applied to the brain.

We conclude that this theory is an interesting approach to the mind-matter relation which might complement more neuroscientific approaches such as the Global Neuronal Workspace model. Depending on whether a phenomenological analysis confirms that there are qualia which cannot be distinguished by mere reference to communicable relations, the model may or may not have to be extended in some form or the other to talk about the hard problem of consciousness. Since it describes the physical as closed (Section 3.6), like IIT, it also might have to address the problems mentioned in Remark A.1 and Appendix A.2.

5. Application

In this section, we carry out a first application of the phenomenological grounding. The application does not refer to a particular model of consciousness, but is of general nature: General assumptions, which one might want to impose when constructing models of consciousness, have strong implications for the physical domain.

The results of this section can be taken as an example to show that the mathematical structure associated to label spaces $L$ in a model of consciousness, can, if the model does not postulate the physical as closed, have observable implications for the dynamics of the physical domain. These implications may be quite specific even if the laws $\mathcal{L}$ of a model have not yet been specified.

“It is difficult to overstate the significance of the concept of symmetry in its many guises to the development of modern physics.” [Mar03]
5.1. Novel Symmetries. In Section 3.4 we have argued that a necessary condition for a model of consciousness to be well-defined is that it carries a symmetry $(\text{Aut}(L)^k, \sigma)$. In physics, symmetries have played a pivotal role in the discovery of new theories, in the prediction of new experimental results and in understanding general properties of nature, e.g. via Noether’s theorem. Thus the question arises of whether the existence of the symmetry $(\text{Aut}(L)^k, \sigma)$ similarly has significant implications in the framework considered here.

In this section, we start to answer this question by investigating the relation between the symmetry $(\text{Aut}(L)^k, \sigma)$ of a model of consciousness and known symmetries of physical theories. To this end, instead of considering specific groups and their action on specific physical theories, we consider the class of all symmetries which a particular physical theory $T_P$ can carry.

**Definition 5.1.** We denote by $S(T_P)$ all symmetries of a physical theory $T_P$, i.e. all pairs $(G, \phi)$ of a group $G$ and corresponding action $\phi$ which satisfy Definition 3.23 with respect to $T_P$.

The general idea is this: By definition, any model of consciousness $M$ is based on and extends a particular physical theory $T_P$. Furthermore, it carries a symmetry $(\text{Aut}(L)^k, \sigma)$. The latter contains an action $\tilde{\sigma}$ which operates on the physical trajectories. If this action $\tilde{\sigma}$ satisfies all defining criteria of a symmetry of $T_P$ that is to say, if it is itself a symmetry of $T_P$, then the symmetry $(\text{Aut}(L)^k, \tilde{\sigma})$ can be considered ‘known’: Its existence is deductible from $T_P$ and in fact in almost all cases the symmetry will have been identified by those parts of natural science which work with the theory $T_P$.

If, however, $(\text{Aut}(L)^k, \tilde{\sigma})$ does not constitute a symmetry of $T_P$ according to Definition 3.23 there is no way one could have identified $(\text{Aut}(L)^k, \tilde{\sigma})$ as a symmetry of some sort based on an analysis of $T_P$. In this case, $(\text{Aut}(L)^k, \sigma)$ constitutes a novel symmetry whose implications have not been analysed by the part of natural science dedicated to the study of $T_P$. In so far as a model $M$ describes reality to some extend (just as $T_P$, but taking into account consciousness), one might say that $(\text{Aut}(L)^k, \sigma)$ constitutes a symmetry of reality which one cannot see if one looks at reality through $T_P$ alone.

The key defining property of any symmetry of $T_P$ is that the corresponding action leaves the set $D_P$ of solutions of $T_P$ invariant. The action $\tilde{\sigma}$, on the other hand, leaves the set $D|_P$ invariant, the physical trajectories which are part of solutions $M$.

Therefore, if those two sets agree, $D|_P = D_P$, the action $\tilde{\sigma}$ has almost all that is needed for it to constitute a symmetry of $T_P$. The last equality, however, is precisely the defining property of the closure of the physical (Definition 3.6). Therefore, there is a relation between whether a model $M$ describes the physical as closed and whether its symmetry $(\text{Aut}(L)^k, \sigma)$ is is a novel, previously unconsidered symmetry of the theory $T_P$ on which the model is based.

The goal of this section is to make this relation between the closure of the physical and the novelty of the symmetry $(\text{Aut}(L)^k, \tilde{\sigma})$ precise. To this end we prove three

---

\[38\text{Cf. Definition 3.23 and equation (3.23) and note that without loss of much generality, we have assumed } K|_P = K_P \text{ after (3.17).}\]

\[39\text{It might be the case that } (\text{Aut}(L)^k, \tilde{\sigma}) \text{ is a symmetry of an improved or extended physical theory } T_P. - \text{Cf. Footnote 12.}\]

\[40\text{Cf. Equation (3.24). The set } D|_P \text{ is defined in (3.10).}\]
lemmas below:

- Lemma 5.2 shows that if $M$ describes the physical as closed, it follows that $(\text{Aut}(L)^k, \tilde{\sigma})$ is basically a symmetry of $T_P$. Here, ‘basically’ refers to the fact that an action needs to satisfy effectivity in order to constitute a symmetry (cf. Footnote 28). Since effectivity might (and generally will) get lost when passing from $\sigma$ to $\tilde{\sigma}$, $(\text{Aut}(L)^k, \tilde{\sigma})$ is not quite a symmetry of $T_P$. However, Lemma 5.2 shows that the difference between $(\text{Aut}(L)^k, \tilde{\sigma})$ and a symmetry of $T_P$ is given by a group-action which is trivial.

- Lemma 5.3 is intended to prove the opposite case. It shows that if $M$ does not postulate the physical as closed, and if $\sigma$ and the solutions of $M$ satisfy two general properties, then $(\text{Aut}(L)^k, \tilde{\sigma})$ is a novel, previously unknown symmetry as judged based on $T_P$. The assumptions used in this lemma are, however, not fully convincing.

- Lemma 5.4 has the same goal as Lemma 5.3 but uses a different and more plausible general assumption. It shows that if the solutions of $M$ and the initial conditions of $T_P$ are not in a 1:1 correspondence (which implies that $M$ does not postulate the physical as closed), then $(\text{Aut}(L)^k, \tilde{\sigma})$ is a novel, previously unknown symmetry as judged based on $T_P$.

The last two lemmas can be interpreted as showing that in general, if $M$ does not postulate the physical as closed, $(\text{Aut}(L)^k, \tilde{\sigma})$ will constitute a novel symmetry with respect to $T_P$. Further assumptions are necessary to prove this because the definition of symmetries does not strongly constrain how the symmetry group acts outside of the set of solutions. We note that in both lemmas, essential properties of a symmetry are violated (invariance of $D_P$ in Lemma 5.3 and compatibility of the group-action with the group multiplication in Lemma 5.4).

**Lemma 5.2.** If $M$ describes the physical as closed, $(\text{Aut}(L)^k, \tilde{\sigma})$ is an extension of a symmetry of $T_P$ by a group $N$ which acts trivially on $K_P$.

**Proof.** Since $(\text{Aut}(L)^k, \sigma)$ is a symmetry of $M$, by Definition 3.23, it leaves the set $D$ of solutions invariant. By (5.22), this implies that $\tilde{\sigma}$ leaves the set $D|_P$ invariant. The model $M$ describes the physical as closed if and only if $D|_P = D_P$. Thus the action $\tilde{\sigma}$ leaves $D_P$ invariant, which shows that condition (b) of Definition 5.23 is satisfied with respect to $T_P$.

The action $\tilde{\sigma}$ of $\text{Aut}(L)^k$ on $K_P$ might nevertheless not constitute a symmetry of $T_P$ because the action might not be effective. It turns out, however, that the difference between a symmetry of $T_P$ and $(\text{Aut}(L)^k, \sigma)$ is merely given by a subgroup of $\text{Aut}(L)^k$ which acts trivially on $K_P$. This subgroup is defined as

$$N := \{ s \in \text{Aut}(L)^k \mid \tilde{\sigma}_s(k) = k \text{ for all } k \in K_P \} .$$

(5.1)

This is a normal subgroup of $\text{Aut}(L)^k$, because for $s \in N$, $g \in \text{Aut}(L)^k$ and all $k \in K_P$, $\tilde{\sigma}_{gs^{-1}}(k) = k$. Hence we can build the quotient group $\text{Aut}(L)^k/N$ which corresponds to

$$\text{Aut}(L)^k/N \times K_P \to K_P$$

$$\tilde{sN}.k \to \tilde{\sigma}_s(k) ,$$

(5.2)

where $sN$ denotes the coset of $s$, i.e. an element of $\text{Aut}(L)^k/N$. This action is effective because every $s \in \text{Aut}(L)^k$ which fixes all $k \in K_P$ by definition of $N$ is a representative

---

**MODELS OF CONSCIOUSNESS 43**
of the coset $e \mathcal{N}$, i.e. a representative of the neutral element of $\text{Aut}(L)^k/\mathcal{N}$. Therefore, $\text{Aut}(L)^k/\mathcal{N}$ with the action (5.2) is a symmetry of $T_P$. Put differently, $\text{Aut}(L)^k$ is an extension of a symmetry of $T_P$ by $\mathcal{N}$. By definition of $\mathcal{N}$ in (5.1), $\tilde{\sigma}|_{\mathcal{N}}$ is a $\mathcal{N}$-action on $P$ which acts trivially on $K_P$ (it does not change a single trajectory). \hfill \Box

**Lemma 5.3.** If $M$ does not postulate the physical as closed, if $\tilde{\sigma}$ acts transitively on $\mathcal{D}|_{P}$, and if some, not all, solutions of the model $M$ contain physical trajectories which are solutions of the physical theory $T_P$, then $(\text{Aut}(L)^k, \tilde{\sigma})$ is not a symmetry of $T_P$.

**Proof.** Since $(\text{Aut}(L)^k, \sigma)$ is a symmetry of $M$, by (3.22), the action $\tilde{\sigma}$ leaves $\mathcal{D}|_{P}$ invariant. Since $M$ does not postulate the physical as closed, we have $\mathcal{D}|_{P} \neq \mathcal{D}_P$. However, this does not exclude the possibility that $\tilde{\sigma}$ leaves both $\mathcal{D}|_{P}$ and $\mathcal{D}_P$ invariant.

The third premise of the lemma states that $\mathcal{D}|_{P} \cap \mathcal{D}_P \neq \emptyset$ and that $\mathcal{D}|_{P}$ is not a proper subset of $\mathcal{D}_P$. Therefore, there are trajectories $k \in \mathcal{D}|_{P} \cap \mathcal{D}_P$ and $k' \in \mathcal{D}|_{P} \setminus \mathcal{D}_P$. Since $\tilde{\sigma}$ acts transitively, there is a $\tilde{s} \in \text{Aut}(L)^k$ such that

$$\tilde{\sigma}_{\tilde{s}}(k) = k'.$$

Therefore, $\tilde{\sigma}$ does not leave $\mathcal{D}_P$ invariant, which implies that $(\text{Aut}(L)^k, \sigma)$ cannot be a symmetry of $T_P$. \hfill \Box

The second and third premises of Lemma 5.3 are not strongly convincing from a conceptual point of view. Whereas the third premise could be argued for as a particular kind of non-closure of the physical, the second premise seems hard to justify in general: There does not seem to be a good reason of why the action $\tilde{\sigma}$ should act transitively on $\mathcal{D}|_{P}$. To this end, we provide another lemma which takes a different route, using the fact that the laws of many physical theories $T_P$ are such that elements of the state space $P$ determine solutions $\mathcal{D}_P$ uniquely. In fact, in Section 3.3 we have assumed this to be the case for the physical theories considered here.\footnote{This assumption is only relevant at this point, it has not been used before and hence could also be dropped.}

Denote the restriction of a trajectory $(p_t)_{t \in I} \in \mathcal{D}_P$ of the physical theory $T_P$ to an arbitrary but fixed $t_0 \in I$ by

$$\text{res}_{T_P} : \mathcal{D}_P \rightarrow P. \quad (5.3)$$

By the assumption just mentioned, this map has an inverse $\text{res}_{T_P}^{-1}$: For any $p \in P$, the laws of $T_P$ determine a unique trajectory $\text{res}_{T_P}^{-1}(p)$ such that $\text{res}_{T_P}(\text{res}_{T_P}^{-1}(p)) = p$ for any $p \in P$ and $\text{res}_{T_P}^{-1}(\text{res}_{T_P}(k)) = k$ for any $k \in \mathcal{D}_P$.

Similarly, we may define a restriction map $\mathcal{D} \rightarrow P$ for the model $M$ by mapping a trajectory $(\hat{t}, p_t) \in \mathcal{D}$ to $p_{t_0} \in P$. Restricting this map to $\mathcal{D}|_{P}$ gives the analogon of (5.3),

$$\text{res}_{M} : \mathcal{D}|_{P} \rightarrow P. \quad (5.4)$$

Whereas this map may well be surjective, in contrast to (5.3), it will in general not be injective: There will typically be several trajectories in $\mathcal{D}$ whose $p_{t_0}$-value is the same. The following lemma shows that this last requirement is sufficient to ensure that $(\text{Aut}(L)^k, \tilde{\sigma})$ is not a symmetry of $T_P$.

**Lemma 5.4.** If $M$ does not postulate the physical as closed and if (5.4) is not injective, $(\text{Aut}(L)^k, \tilde{\sigma})$ is not a symmetry of $T_P$. 

Proof. Since (Aut$$(L)^k, \sigma$) is a symmetry of $M$, by (3.22), the action $\tilde{\sigma}$ leaves $\mathcal{D}|_{P}$ invariant. Since $M$ does not postulate the physical as closed, we have $\mathcal{D}|_{P} \neq \mathcal{D}_P$. However, this does not exclude the possibility that $\tilde{\sigma}$ leaves both $\mathcal{D}|_{P}$ and $\mathcal{D}_P$ invariant.

The third premise of the lemma states that $\mathcal{D}|_{P} \cap \mathcal{D}_P \neq \emptyset$ and that $\mathcal{D}|_{P}$ is not a proper subset of $\mathcal{D}_P$. Therefore, there are trajectories $k \in \mathcal{D}|_{P} \cap \mathcal{D}_P$ and $k' \in \mathcal{D}|_{P} \setminus \mathcal{D}_P$. Since $\tilde{\sigma}$ acts transitively, there is a $\tilde{s} \in \text{Aut}(L)^k$ such that

$$\tilde{\sigma}_{\tilde{s}}(k) = k'.$$

Therefore, $\tilde{\sigma}$ does not leave $\mathcal{D}_P$ invariant, which implies that $(\text{Aut}(L)^k, \sigma)$ cannot be a symmetry of $T_P$. \hfill \Box
Note that it is sufficient for there to be one parameter $t_0 \in I$ for which the corresponding map (5.4) is not injective.

*Proof.* Let $(G, \varphi)$ be a symmetry of $TP$. The restriction map $\phi_k$ determines how $\varphi$ acts on $TP$: It maps $\text{res}_{TP}(k)$ to $\text{res}_{TP}(\varphi_g(k))$, i.e. it acts as

$$\varphi^\circ : G \times P \to P$$

$$(g, p) \mapsto \varphi^\circ_g(p) := \text{res}_{TP} \circ \varphi_g \circ \text{res}_{TP}^{-1}(p),$$

(5.5)

where $\circ$ denotes function composition. Since (5.4) is an isomorphism of sets, it follows that (5.5) is a $G$-action on $P$. Furthermore, it follows that

$$\varphi_g(k) = \text{res}_{TP}^{-1} \circ \varphi^\circ_g \circ \text{res}_{TP}(k)$$

(5.6)

needs to hold for all $k \in D_P$.

Consider now the action $\tilde{\sigma}$ which is part of the symmetry $(\text{Aut}(L)^k, \sigma)$ of $M$. Since $\tilde{\sigma}$ leaves $D|p$ invariant, we can investigate how $\tilde{\sigma}$ operates on $P$ by using (5.4). Since $\text{res}_M$ is not injective, there is no inverse map $\text{res}_M^{-1} : P \to D|p$. However, we can still consider the pre-image of any point $p \in P$, which we will denote as $\text{res}_M^{-1}(p)$; it consists of all those trajectories in $D|p$ which restrict to the same point $p \in P$. Because $\text{res}_M$ is not injective, there are $p \in P$ such that $\text{res}_M^{-1}(p)$ has cardinality larger one.

Let $g \in \text{Aut}(L)^k$. If $g$ acts on $D|p$ via $\tilde{\sigma}$, any $p \in P$ will be mapped to one of the points in the set $\text{res}_{M}(\tilde{\sigma}_g(\text{res}_M^{-1}(p)))$, where we have used the usual notation $f(A) = \{f(x) | x \in A\}$ to denote images of sets $A$ with respect to functions $f$. Therefore, we can describe how $\tilde{\sigma}$ transforms states $p \in P$ introducing a map

$$\tilde{\sigma}_g^0(p) := \text{res}_M \circ \tilde{\sigma}_g \circ \text{res}_M^{-1}(p),$$

(5.7)

which, in contrast to (5.5), is not a function but rather a correspondence: It maps states $p \in P$ to sets $\tilde{\sigma}_g^0(p)$ of states.

We have shown above that any symmetry of $TP$ necessarily satisfies (5.6). Thus, if $\tilde{\sigma}$ is a symmetry of $TP$, it needs to satisfy

$$\tilde{\sigma}_g(k) = \text{res}_{TP}^{-1} \circ \tilde{\sigma}_g^0 \circ \text{res}_{TP}(k)$$

for all $k \in D_P$, which together with (5.7) implies

$$\tilde{\sigma}_g(k) = \text{res}_{TP}^{-1} \circ \text{res}_M \circ \tilde{\sigma}_g \circ \text{res}_M^{-1} \circ \text{res}_{TP}(k).$$

(5.8)

This, however, is a contradiction to $\tilde{\sigma}$ being a group action, as we now show.

If $M$ does not postulate the physical as closed, (5.8) is a strong restriction on how $\tilde{\sigma}$ acts: It determines its action on $D_P$ in terms of its action on $D|p$. This is problematic because due to the non-injectivity of $\text{res}_M$, the composition $\text{res}_M^{-1} \circ \text{res}_{TP}$ is not an isomorphism, so that many properties of a group action might be violated. However, it turns out that at this level of generality, only the compatibility property, $\tilde{\sigma}_{g \cdot h} = \tilde{\sigma}_g \circ \tilde{\sigma}_h$ for all $g, h \in \text{Aut}(L)^k$, is necessarily violated. (Where $g \cdot h$ denotes the group multiplication of $g$ and $h$ in $\text{Aut}(L)^k$.)

Let $k \in D_P$. In order to calculate $\tilde{\sigma}_g \circ \tilde{\sigma}_g(k)$ according to (5.8), we need to assume that $\tilde{\sigma}_g(k) \in D_P$. If this is not the case, the set $D_P$ is not invariant w.r.t. $\tilde{\sigma}$, which implies that $(\text{Aut}(L)^k, \tilde{\sigma})$ is not a symmetry of $TP$ as claimed. If this is the case,
according to (5.8), we have
\[ \tilde{\sigma}_g \circ \tilde{\sigma}_g(k) = \operatorname{res}_{T_P}^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_g \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_{T_P} \circ \operatorname{res}_M \circ \tilde{\sigma}_h \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_{T_P}(k) \]
\[ = \operatorname{res}_{T_P}^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_g \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_h \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_{T_P}(k). \]
Since \( \operatorname{res}_M \) is not injective, \( \operatorname{res}_M^{-1} \circ \operatorname{res}_M \) is a correspondence rather than a function. However, \( \tilde{\sigma}_g, \tilde{\sigma}_h \) and \( \tilde{\sigma}_{g,h} \) are bijections on \( \mathcal{K}_P \), which implies that \( \tilde{\sigma}_g \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_h \) is a correspondence as well and not a function. Therefore, \( \tilde{\sigma}_g \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_h \) cannot be equal to \( \tilde{\sigma}_g \circ \tilde{\sigma}_h = \tilde{\sigma}_{g,h} \), which implies that the right hand side of the last equation cannot be equal \( \operatorname{res}_{T_P}^{-1} \circ \operatorname{res}_M \circ \tilde{\sigma}_g \circ \operatorname{res}_M^{-1} \circ \operatorname{res}_{T_P}(k) \). Using again (5.8), in total we have shown that (5.8) implies
\[ \tilde{\sigma}_g \circ \tilde{\sigma}_g(k) \neq \tilde{\sigma}_{g,h}(k), \]
which is a contradiction to \( \tilde{\sigma} \) being a group action, as claimed. \( \square \)

Note that in Lemmas 5.3 and 5.4 the assumption of \( M \) not postulating the physical as closed is in fact implied by the third, respectively second, assumption. Whereas it therefore would not have been necessary to state this assumption explicitly when formulating the lemmas, we have chosen to include it for the sake of clarity for those who have a quick read.

The general upshot of this section is that whether (\( \operatorname{Aut}(L)^k, \sigma \)) is a novel, previously unknown symmetry (as judged based on \( T_P \)) depends not so much on which physical theory \( T_P \) one chooses to build a model of consciousness, but rather on whether one considers models which postulate the physical as closed or not\(^{42}\). In the next section, we will see that this can have far-reaching empirical consequences.

5.2. Qualia Determined Locally. Models of consciousness \( M \) are based on and extend physical theories \( T_P \). Therefore, it is reasonable to try to propose laws for \( M \) which contain the laws of \( T_P \) as a limiting or special case so as to make sure that in most physical laboratory experiments, the predictions of \( T_P \) still hold true\(^{43}\). In this section, we show that if the laws of \( M \) being proposed contain a simple term which is ubiquitous among fundamental physical theories, and if \( M \) furthermore satisfies a general condition which expresses the idea that qualia are determined locally rather than globally in space, this has far reaching empirical consequences. Here, the space in question is ‘physical space’ – the complement of time, the “three-dimensional extent in which objects and events occur and have relative position and direction” [EB98] – and the terms ‘global’ and ‘local’ refer roughly to whether the physics in all of space play a role in determining qualia, or merely a subset thereof.

We will use the symbol \( \Sigma \) to denote physical space. The simple term in question is a derivative of the physical states \( p_t \) with respect to coordinates \( (x^\mu)_{\mu=1,2,3} \) of \( \Sigma \), which we call ‘spatial derivative’ and which we denote as
\[ \partial_\mu := \frac{\partial}{\partial x^\mu}. \quad (5.9) \]
In order to spell out the assumptions in detail, we need to fix precisely what we mean by ‘space’, what is implied by assuming a quantity to ‘depend on space’ and how to define spatial derivatives of such quantities. This will be done in the next subsection. We remark that the following choice is not universal and other choices are possible, e.g. if one wishes to describe operator-valued distributions which appear in quantum field theory. However, the following choice seems to be sufficiently general to make the point while still remaining largely accessible. Adaptations to other formalizations of these three notions should be straightforward. Also, we remark that the reader could just jump over this mathematical setup as most of what follows is comprehensible with an intuitive understanding of these three terms.

5.2.1. Mathematical Setup. We take physical space \( \Sigma \) to be a smooth Riemannian 3-manifold. A quantity \( \tilde{p} \) will be called ‘dependent on space’ if it is a smooth map \( \tilde{p} : \Sigma \to E \) from \( \Sigma \) to a smooth manifold \( E \). This includes the case of simple physical fields or wave-functions, where \( E \) is \( \mathbb{R}^n \) or \( \mathbb{C}^n \) as well as the case of general physical fields, where \( E \) is a smooth vector bundle over \( \Sigma \) and \( \tilde{p} \) a section thereof, e.g. the associated bundle of some principal bundle in the case of gauge theories. We denote the set of maps from \( E \) to itself by \( C(E, E) \) and use \( C^0(E, E) \) and \( C^\infty(E, E) \) to indicate continuous and smooth subsets thereof, respectively.

We denote coordinate charts of \( \Sigma \) by \( (U, \varphi^\Sigma) \) and its component functions (local coordinates) by \( (x^\mu)_{\mu=1,2,3} \). Coordinate charts of \( E \) will be denoted by \( (V, \varphi^E) \) with component functions \( (y^a)_{a=1,...,n} \). Composing the map \( \tilde{p} \) with \( \varphi^E \) yields the component functions \( \tilde{p}^a \), with respect to which we may define the spatial derivative (5.9) as

\[
\frac{\partial}{\partial x^\mu} \tilde{p}^a(x) .
\] (5.10)

If \( E \) does not have a linear structure, these coordinate expressions will in general not have geometric meaning. Rather, they will constitute one summand in a coordinate expression of coordinate invariant generalizations of (5.9) such as a connection in case \( E \) is a vector bundle, or a covariant (‘gauge invariant’) derivative in case \( E \) is associated to a principal bundle. Nevertheless, the coordinate expression (5.10) will be sufficient for the purpose of this section.

5.2.2. First Assumption. Having fixed the mathematical setup, we proceed to specify the first assumption of Proposition 5.10. To this end, we note that the laws of all fundamental physical theories contain spatial derivatives, and except for quantum field theory, the notion of spatial derivative which they employ is compatible with the formulation we have just introduced. Put differently, the state space \( P^T \) of most fundamental physical theories \( T_P \) contains quantities \( \tilde{p} \) which are dependent on space in the sense introduced above, and the laws of these theories contain the term (5.10). If one wishes to build a model of consciousness \( M \) based on such a fundamental physical theory \( T_P \), the dynamical variables of the model \( M \) will by definition have to include the quantities \( \tilde{p} \). If one furthermore wishes to preserve the laws of \( T_P \) to some extend when constructing the model \( M \), one might want to include the derivative terms (5.10) into the laws of \( M \). We summarise this as follows:

\[44\text{As indicated at the beginning of this section, the laws of } M \text{ could be chosen so as to yield the terms (5.10) only after a limiting procedure. The present formalization of the ideas underlying this section does not cover this case. However, the adaptation to this case seems relatively straightforward.}\]
Definition 5.5. A model of consciousness $M$ inherits spatial derivatives if its dynamical variables $d$ or background structure $b$ contain $\Sigma$, if the physical state space $P$ comprises quantities $\tilde{p}$ which are dependent on space (defined above) and if furthermore the laws $L$ of the model contain equations which in turn contain the spatial derivative $(5.10)$ if written in terms of coordinates of $\Sigma$ and $E$.

Since the laws $L$ are applied to kinematically possible trajectories $K$, the spatial derivative in $(5.10)$ will in fact act on elements $\tilde{p}_t$ of a corresponding family of states. We abbreviate $(5.10)$ in this case as $\partial_{\mu}\tilde{p}^\alpha_t(x)$.

5.2.3. Second Assumption. The second assumption of Proposition 5.10 expresses the idea, mentioned above, that qualia are determined locally, rather than globally, in space. This idea can actually be fleshed out in very simple terms without invoking further mathematical structures beyond what we have used so far. To this end, we consider two solutions $(\tilde{\ell}_t, \tilde{p}_t)_{t\in T}$ and $(\tilde{\ell}_t', \tilde{p}_t')_{t\in T}$ of the model whose label states agree for all parameters $t \in T$ except for a small subset $T_\delta$, where they disagree. The following definition amounts to postulating that a difference of this sort in qualia comes about due to a local, rather than global, difference in the physical states of the solutions. A detailed explanation follows below.

Definition 5.6. A model of consciousness $M$ which inherits spatial derivatives determines qualia locally if for any two solutions $(\tilde{\ell}_t, \tilde{p}_t)_{t\in T}$ and $(\tilde{\ell}_t', \tilde{p}_t')_{t\in T} \in D$ which satisfy

$$ \tilde{\ell}_t \neq \tilde{\ell}_t' \text{ on some small } T_\delta \subseteq T \text{ but } \tilde{\ell}_t = \tilde{\ell}_t' \text{ otherwise,} $$

there is a $t \in T_\delta$ and open $U \subseteq \Sigma$ such that $\tilde{p}_t|U \neq \tilde{p}_t'|U$ but $\tilde{p}_t'|\Sigma \setminus U = \tilde{p}_t|\Sigma \setminus U$.

Equation $(5.11)$ expresses a pointwise condition: The requirement is that for all $t \in T \setminus T_\delta$, we have $\tilde{\ell}_t = \tilde{\ell}_t'$, and for all $t \in T_\delta$, we have $\tilde{\ell}_t \neq \tilde{\ell}_t'$. The latter is satisfied if there is at least one $i \in \{1, \ldots, k\}$ such that $\ell_t^i \neq \ell_t'^i$, i.e. if the qualia of at least one experiencing subject is different for all $t \in T_\delta$. If there is a small (local) difference of qualia of this sort (“in time”), Definition 5.6 requires there to be a local difference of the physical states in space. Here, the notion of locality which is employed is very weak: The difference of two physical states is local in space if the physical states agree outside of some proper open subset of $\Sigma$, i.e. if they are not completely different. What is essential in Definition 5.6 is that the difference between the physical states, if it exists, is not constant in space.

5.2.4. Third Assumption. Finally, the last ingredient necessary to state the main result of this section is the invariance of laws of a theory under symmetry transformations, as we now explain.

In Section 3.3 we have described the general picture of what constitutes a scientific theory $T$ according to contemporary philosophy of science. According to this picture, a theory contains some laws which select from the set of all kinematically possible trajectories $K$ a set of solutions $D$. These laws may – and generally will – contain given that a regularization of the spatial derivative can be assumed to still satisfy properties which are similar (up to small error terms) to the axioms of a connection.

We assume a suitable notion of ‘small subset’ to be given in an application of the theory. Formally, Proposition 5.10 is true if every proper subset of $T$ is considered a ‘small subset’. However, in practise one might want to be more restrictive in which subsets of $T$ to define as small because this will make it easier for Definition 5.6 to be satisfied. The notion of smallness can be formalized here by defining it in terms of a subset $s[T] \subset P(T)$ of the power set of $T$.
equations. We have deliberately not made any assumptions about the mathematical formalism which is being used in the statement of the equations, for a) there are many mathematical formalisms in use in science and it seems premature to pick any one as best suited for the scientific study of consciousness, and, more importantly, because b) it seems very plausible that novel developments in mathematics will prove to be crucial in formulating models of consciousness. Thus, in order to specify what we mean by invariance, we need to consider an equation formulated in a general formal theory\(^{46}\) which we denote by \(\varepsilon\). In our case, where as part of the laws of a theory, the equation \(\varepsilon\) under consideration is a selection criterion on kinematically possible trajectories \(K\), the equation contains special variable symbols which need to be replaced, according to a set of rules, by terms which constitute specific trajectories \(k \in K\). We use the notation \(\varepsilon[k]\) to denote the equation which one obtains once a trajectory \(k \in K\) has been “plugged into” \(\varepsilon\). Using this notation, invariance of equations of a theory \(T\) with respect to symmetries may be defined as follows. Let \((G, \phi)\) be a symmetry of \(T\).

**Definition 5.7.** An equation \(\varepsilon\) of \(T\) is \(G\)-invariant if for every \(k \in D\) and every \(g \in G\), the equation \(\varepsilon[k]\) is equivalent, according to the formal theory used to define \(\varepsilon\), to the equation \(\varepsilon[\phi_g(k)]\). I.e., if the equation \(\varepsilon[k]\) can be derived, according to the deduction rules and axioms of the formal theory, from the equation \(\varepsilon[\phi_g(k)]\) and vice versa. An equation is invariant if it is \(G\)-invariant for every symmetry \((G, \phi)\) of \(T\). Finally, a theory \(T\) has invariant laws if the equations contained in its laws are invariant.

Invariance is an important concept in fundamental physics. Laws of fundamental physical theories \(T_P\) are generally required to be \(G\)-invariant with respect to known symmetries \((G, \phi)\) of \(T_P\), and the requirement of invariance has been the basis of novel predictions and discoveries in physics. Most notably, in the context of gauge theories, the requirement of invariance has been used to predict the existence of new physical fields. E.g., “through demanding a local gauge (or isotopic spin, or \(SU(2)\)-phase) invariance, [Yang and Mills] were led to introduce a new field, an \(SU(2)\) gauge field. (...) In the following decade, detailed predictions of, for example, the existence and properties of the weak neutral current and the \(Z_0\) and \(W_\pm\) bosons required by the gauge invariance were confirmed with amazing accuracy.” [Mar03, p. 39f.] Similarly, the invariance under a local \(SU(3)\) symmetry led to the introduction of gluon

---

\(^{46}\)Here, we use the term ‘formal theory’ in the sense of mathematical logic. Thus, a formal theory consists of a set of symbols, which allows to define terms and formulas, a set of deduction rules, which allow to prove that a formula is a consequence of another, and a choice of sentences as axioms \([EFT84]\). We denote the set of symbols by \(S\). In the following footnotes, we explain some of the concepts used in the main text in terms of formal theories. For the sake of clarity, we will sometimes use the term ‘mathematical formalism’ synonymous with ‘formal theory’.

\(^{47}\)According to mathematical logic \([EFT84]\), equation is a formula \(t_L = t_R\), where \(t_L\) and \(t_R\) are terms of \(S\). Terms are defined recursively to be variables, constants or to be of the form \(f t_0 ... t_{n-1}\), where \(f\) is an \(n\)-ary function symbol of \(S\) and \(t_0, ..., t_{n-1}\) are terms. As indicated in the main text, we use the symbol \(\varepsilon[k]\) to denote the equation \(t_L[k] = t_R[k]\) which has been obtained from \(t_L = t_R\) by replacing special variables of \(S\) with terms which constitute specific trajectories \(k \in K\) according to some rules. Only after this replacement can we speak about \(\varepsilon[k]\) being true or not: \(\varepsilon[k]\) is true if and only if \(k\) is a solution of the model. As an example for such a replacement, consider the term \(5X10\) which is contained in an equation \(\varepsilon\) of the laws of a theory if this theory contains spatial derivative. One needs to replace the variable \(\tilde{p}\) by the space-dependent part \(\tilde{p}_t\) of the physical state \(p_t\) of a particular trajectory \(k = (\tilde{t}, p_t)_{\tilde{t} \in Z}\), among other things, in order to be able to evaluate whether the equation holds for the trajectory under consideration.
fields and a “successful theory of the strong interactions, quantum chromodynamics (QCD)” \[Mar03\, p. 41\]. This shift from “taking symmetries to be interesting but after-the-fact features of some known law(s)” to “starting from symmetries and then deducing from these various physical consequences, laws, etc.” \[Mar03\, p. 32\] is already apparent in Einstein’s work on special relativity theory \[Mar03\].

To a first approximation, the logic of these discoveries can be reconstructed as follows: At a particular time, some physical theory \(T_P\) was known, e.g. Quantum Electrodynamics in the case of Yang and Mills. For various reasons, among them e.g. a “failure of the standard methods for constructing theories of the new[ly discovered] particles and their interactions” \[Mar03\], it was postulated that the theory \(T_P\) carries a previously unconsidered symmetry (\(G, \phi\)). Due to an invariance requirement on the laws of \(T_P\), the laws had to be changed so as to establish \(G\)-invariance, and the specific nature of the postulated symmetry necessitated the introduction of new physical fields to this end.

In Section 5.1, we have found a somewhat similar situation with respect to models of consciousness: We have seen that compared to the physical theory \(T_P\) based on which a model \(M\) is constructed, the symmetry (\(\text{Aut}(L)^k, \sigma\)) constitutes a novel symmetry if, given an additional assumption, the model \(M\) does not postulate the physical as closed. In what follows, we prove that if Definitions 5.5 and 5.6 also hold, invariance of the model \(M\) with respect to (\(\text{Aut}(L)^k, \sigma\)) can only be satisfied if there is a novel physical field, just as in the gauge theory cases of physics.

Definition 5.5 gives us some handle on the mathematical formalism being used in an equation \(e\) of \(M\): The formal theory needs to be such that the spatial derivatives (5.10) are well-defined. Unfortunately, this is not enough information to deduce the standard invariance requirement of spatial derivatives used in physics from Definition 5.7. To this end, we explicitly state this requirement in the following definition. Further assumptions on the mathematical formalism may allow a deduction of Definition 5.8 from Definition 5.7. In the following definition, \(t\) denotes spatial derivative terms of the formal theory, e.g. 5.10.

**Definition 5.8.** A spatial derivative term \(t\) is invariant, if for every symmetry (\(G, \phi\)) of a theory \(T\) and every \(g \in G\), \(t[k]\) and \(t[\phi_g(k)]\) have the same additive structure.

The upshot of Section 5.2.4 is that invariance in the form of Definition 5.8 is both a standard requirement on physical theories and a reasonable assumption. In fact, if a theory contains non-invariant spatial derivative terms, strong conceptual and mathematical problems will generally appear.

**5.2.5. Proposition.** We now combine the last three definitions with the symmetry (3.22) that any model of consciousness needs to posses. The connection with Section 5.1 will be established in the subsequent corollary.

To see why it is possible to make this connection, note that the automorphism group \(\text{Aut}(L)\) consists of all bijective homomorphisms of the space \(L\). Thus, most elements \(s \in \text{Aut}(L)\) will only transform a subspace \(L \subset L\) while keeping the complement

\[48\]In terms of a formal theory as introduced in Footnotes 16 and 17, the requirement is that the function symbol ‘+’, which needs to be contained in the formal theory under consideration if a model contains spatial derivatives (Definition 5.8), appears in \(t[k]\) and \(t[\phi_g(k)]\) in isomorphic ways: E.g., if \(t[k]\) is of the form \(t_1 + t_2\) for some terms \(t_1\) and \(t_2\) which do not contain the function symbol ‘+’, then \(t[\phi_g(k)]\) is also of the form \(t'_1 + t'_2\) for some terms \(t'_1\) and \(t'_2\) which do not contain the function symbol ‘+’.
fixed, \( s(\ell) = \ell \) for all \( \ell \in L \setminus \bar{L} \). Let \( k := (\bar{\ell}_t, p_t)_{t \in I} \in D \) be any solution of \( M \) and consider the trajectory \((\bar{\ell}_t^i)_{t \in I} \) of this solution in the \( i \)th label space. We denote the image of this curve in \( L \) by \( \text{Im}_i(k) \); it consists of those labels which are occupied by the trajectory \((\bar{\ell}_t^i)_{t} \). If this trajectory varies sufficiently (i.e. if \( \text{Im}_i(k) \) is sufficiently large), there will be a \( s \in \text{Aut}(L) \) which transforms only a proper subset \( \bar{\ell} \) of \( \text{Im}_i(k) \).

Accordingly, \((\ell_t^i)_{t} \) is transformed as

\[
\ell_t^i \mapsto \begin{cases} 
\ell_t^i & \text{for } t \in I \setminus I_{\delta} \\
 s(\ell_t^i) & \text{for } t \in I_{\delta}
\end{cases},
\]

where \( I_{\delta} := \{ t \in I \mid \ell_t \in \bar{\ell} \} \) consists of those parameter values \( t \) for which \( \ell_t \in \bar{\ell} \). If we extended this \( s \in \text{Aut}(L) \) to a transformation \( \bar{s} \in \text{Aut}(L)^k \) by identity automorphisms, \( \bar{s} := (\text{id}, \ldots, s, \ldots, \text{id}) \), the solutions

\[
(\bar{\ell}_t, p_t)_{t \in I} \quad \text{and} \quad (\ell_t^i, p_t^i)_{t \in I} := \sigma_{\bar{s}}((\bar{\ell}_t, p_t)_{t})
\]

satisfy condition (5.11). Here, \( \sigma \) denotes the action (3.22) which constitutes the symmetry \( (\text{Aut}(L)^k, \sigma) \) of \( M \).

In short: If there is one single solution \((\bar{\ell}_t, p_t)_{t} \in D \) which contains one single trajectory \((\bar{\ell}_t^i)_{t} \) that varies sufficiently in the above sense, due to the symmetry (3.22), the model necessary contains two solutions (5.13) which satisfy condition (5.11). Due to the general nature of the automorphism group \( \text{Aut}(L) \), this is an extremely weak requirement and we refer to models which do not meet this requirement as trivial models.\(^{49}\)

In a typical model, there will be many such trajectories for various \( \bar{s} \in \text{Aut}(L)^k \).

**Definition 5.9.** A model of consciousness \( M \) is trivial iff there is at least one \( k \in D \) and at least one \( \bar{s} \in \text{Aut}(L)^k \) such that \( k \) and \( k' := \sigma_{\bar{s}}(k) \) satisfy (5.11).

With this we have acquired all terminology to be able to state the main result of this section. Recall that we use \( C(E, E) \) to denote all maps from \( E \) to \( E \).

**Proposition 5.10.** If a non-trivial model of consciousness inherits invariant spatial derivatives and determines qualia locally, the model necessarily contains an object

\[
\Xi_\mu : \Sigma \to C(E, E)
\]

as well as an equation which includes an interaction term between \( \Xi_\mu \) and \( \hat{p} \).

Here, \( \hat{p} \) refers to the space-dependent physical quantities asserted by Definition 5.5.

**Proof.** This proof proceeds in three steps:

- **Step 1:** First, we consider the action of the symmetry \((\text{Aut}(L)^k, \sigma)\), which any model of consciousness needs to carry, on the space-dependent parts of the physical trajectories. As usual, the product rule of the derivative (5.10) yields a term in which the symmetry action is being differentiated.

- **Step 2:** Next, we show that if \( M \) determines qualia locally, this latter term cannot vanish.

\(^{49}\)If ‘small’ subsets are not just proper subsets of \( I \) (cf. Footnote \(^45\)), Condition (5.11) is satisfied provided that \( I_{\delta} \) is small.

\(^{50}\)Due to length restrictions, we cannot enter an analysis of trivial models here. We merely remark that it seems highly questionable whether any trivial model can be put forward which is compatible with the most simple empirical facts about our experience of qualia as defined in Definition 3.8.
Step 3: Finally, we use the invariance requirement to conclude that this term needs to be cancelled by a field (5.14).

Step 1: Let $\bar{s} \in \text{Aut}(L)^k$. The action of $\sigma_s$ on the physical trajectories $\mathcal{K}_\rho$ is given by (5.23). According to Definition 5.5 the physical state space comprises quantities $\bar{p}$ which are dependent on space. Thus any solution $k \in \mathcal{D}$ contains a trajectory $(\bar{p}_t)_{t \in \mathcal{I}}$ of quantities which are dependent on space. We denote the transformation of these trajectories which is induced by $\bar{\sigma}_s$ by

$$
\chi_\bar{s} : (\bar{p}_t)_{t \in \mathcal{I}} \mapsto (\bar{p}_t^\prime)_{t \in \mathcal{I}}.
$$

(5.15)

The $t$-component of this transformation will be denoted by $\chi_\bar{s}^t : \bar{p}_t \mapsto \bar{p}_t^\prime$. According to Definition 5.5 and Section 5.2.1, $\bar{p}_t$ and $\bar{p}_t^\prime$ are smooth maps from $\Sigma$ to a smooth manifold $E$. Thus for fixed $k \in \mathcal{D}$ and $k' := \sigma_s(k) \in \mathcal{D}$, we can represent $\chi_\bar{s}^t(\bar{p}_t)$ as

$$
\chi_\bar{s}^t(x) \bar{p}_t(x),
$$

(5.16)

where $\chi_\bar{s}^t(x)$ is any map $\chi_\bar{s}^t : E \rightarrow E$ which satisfies $\chi_\bar{s}^t(x)[\bar{p}_t(x)] = \bar{p}_t^\prime(x)$. This map quantifies how the space-dependent parts $\bar{p}_t(x)$ and $\bar{p}_t^\prime(x)$ of the two solutions $k$ and $k'$ differ. In case $\bar{p}_t(x) \neq \bar{p}_t^\prime(x)$, any representation necessarily satisfies $\chi_\bar{s}^t(x) \neq \text{id}_E$, where $\text{id}_E$ denotes the identity map on $E$. Furthermore, we may choose a representation such that in case $\bar{p}_t(x) = \bar{p}_t^\prime(x)$ at some $x \in \Sigma$, we have $\chi_\bar{s}^t(x) = \text{id}_E$. Finally, we choose representations $\chi_\bar{s}^t(x)$ to be constant in $x$ if both $\bar{p}_t(x)$ and $\bar{p}_t^\prime(x)$ are constant in $x$.

The last paragraph says that for any choice of $k$ and $\bar{s}$, we may represent $\chi_\bar{s}^t$ by any map $\chi_\bar{s}^t : \Sigma \times E \rightarrow E$ which satisfies $\chi_\bar{s}^t[x, \bar{p}_t(x)] = \bar{p}_t^\prime(x)$ for every $x \in \Sigma$. Since both $\bar{p}_t$ and $\bar{p}_t^\prime$ are smooth in $x$, we may choose the map $\chi_\bar{s}^t$ to be smooth in $x$ as well. In Step 2 of the proof we will use Definition 5.2 to assert a further property of $\chi_\bar{s}^t$. For simplicity, we work with the notation (5.16) in what follows. Furthermore, given a chart $(V_1, \varphi_1^E)$ covering $\bar{p}_t(x)$ and a chart $(V_2, \varphi_2^E)$ covering $\chi_\bar{s}^t(x)\bar{p}_t(x)$, we denote the components of this map as $\chi_\bar{s}^t(x)_a$, where the $a$-index refers to $\varphi_1^E$ and the $b$-index to $\varphi_2^E$. Summation convention will be used as usual, e.g. to express (5.16) as $\bar{p}_t^\prime(x) = \chi_\bar{s}^t(x)_a \bar{p}_t^a(x)$.

We now consider the term (5.10) which is contained in an equation of $M$ according to Definition 5.5. Since we have chosen $\chi_\bar{s}^t(x)$ to be smooth in $x$, we have

$$
\partial_\mu \bar{p}_t^\prime(x) = \partial_\mu (\chi_\bar{s}^t(x)_a \bar{p}_t^a(x)) = \chi_\bar{s}^t(x)_a \partial_\mu \bar{p}_t^a(x) + (\partial_\mu \chi_\bar{s}^t(x)_a) \bar{p}_t^a(x),
$$

(5.17)

which concludes the first step.

Step 2: In general, there is no reason to assume that for arbitrary solutions $k$ and $k' \in \mathcal{D}$, the space-dependent parts $\bar{p}_t(x)$ and $\bar{p}_t^\prime(x)$ of the physical states take the same values in $E$ for all $x \in \Sigma$. However, it might be the case that $\sigma_s$ acts trivially on $\bar{p}_t(x)$ for all $\bar{s} \in \text{Aut}(L)^k$ and all $k \in \mathcal{D}$, which would indeed imply $\bar{p}_t(x) = \chi_\bar{s}^t(x)\bar{p}_t(x)$ for all $x \in \Sigma$ and therefore $\chi_\bar{s}^t(x) = \text{id}_E$ and $\partial_\mu \chi_\bar{s}^t(x)_a = 0$, so that the last term in (5.17) vanishes. However, this cannot be the case for any non-trivial model $M$ which determines qualia locally, as we now explain.

As explained in the paragraph before Proposition 5.10, for a non-trivial model $M$ there is a $k \in \mathcal{D}$ and a $\bar{s} \in \text{Aut}(L)^k$ such that $k$ and $k' := \sigma_s(k)$ satisfy (5.11). We apply the discussion of Step 1 to this $k$ and $\bar{s}$. If $M$ determines qualia locally, according to Definition 5.6 there is a $t \in \mathcal{I}$ and open $U \subseteq \Sigma$, such that $\bar{p}_t|_U \neq \bar{p}_t^\prime|_U$ and $\bar{p}_t|_{\Sigma \setminus U} = \bar{p}_t^\prime|_{\Sigma \setminus U}$. This implies that $\chi_\bar{s}^t(x) = \text{id}_E$ for all $x \in \Sigma \setminus U$. Since $\bar{p}_t|_U \neq \bar{p}_t^\prime|_U$, there is a $x \in U$ such that $\bar{p}_t(x)$ and $\bar{p}_t^\prime(x)$ are not equal, which implies that $\chi_\bar{s}^t(x) \neq \text{id}_E$. Any smooth map between manifolds is continuous [Lee02 Lemma 2.2]. Thus there are
in the equations of the theory $T$. Therefore there is, in general, no term $\Xi(x)$ of equations will, in general, not be invariant with respect to $(\text{Aut}(L)\not\subseteq \Sigma)$ of the physical theory $P$. Indeed, if $(\text{Aut}(L)\not\subseteq \Sigma)$ of the physical theory $P$, then the situation is different if $(\text{Aut}(L)\not\subseteq \Sigma)$ of the physical theory $P$. Two properties of the term $\Xi(x)$ are implied by (5.18). First, the last term of (5.18) takes values in $E$, which implies that the first term needs to take values in $E$ as well. This in turn implies that $\Xi(x) \in C(E,E)$. Second, since in Step 2 we have found that $\partial_x \chi^b(x)^b \not\subseteq \Sigma$ of the physical theory $P$ and if this term and the operator symbol $\partial_x$ appear together in $\epsilon$ as $\partial_x + \Xi(x)$, so that

\[ (\partial_x + \Xi(x)) \chi^b(x)^b \bar{p}^a(x) = \chi^b(x)^b \bar{p}^a(x) \left( \partial_x \chi^b(x)^b \right) \bar{p}^a(x), \tag{5.19} \]

qualifies as an invariant spatial derivative.

Two properties of the term $\Xi(x)$ are implied by (5.18). First, the last term of (5.18) takes values in $E$, which implies that the first term needs to take values in $E$ as well. This in turn implies that $\Xi(x) \in C(E,E)$. Second, since in Step 2 we have found that $\partial_x \chi^b(x)^b \not\subseteq \Sigma$ of the physical theory $P$ and if this term and the operator symbol $\partial_x$ appear together in $\epsilon$ as $\partial_x + \Xi(x)$, so that

\[ (\partial_x + \Xi(x)) \chi^b(x)^b \bar{p}^a(x) = \chi^b(x)^b \bar{p}^a(x) \left( \partial_x \chi^b(x)^b \right) \bar{p}^a(x). \tag{5.19} \]

qualifies as an invariant spatial derivative.

The existence of an object $\Xi(x)$ as in (5.14) per se is not noteworthy. E.g., if $(\text{Aut}(L)\not\subseteq \Sigma)$ is a usual gauge symmetry of the physical theory $T_P$. Proposition 5.10 merely asserts the existence of the usual gauge fields which in (5.19) give rise to the usual gauge-invariant derivative. Indeed, if $(\text{Aut}(L)\not\subseteq \Sigma)$ satisfies the defining properties of a symmetry of $T_P$ and if the spatial derivatives of $T_P$ are assumed invariant (which is generally the case), $\Xi(x)$ is an element of the physical state space $P$, i.e. a known physical field.

However, the situation is different if $(\text{Aut}(L)\not\subseteq \Sigma)$ is not a symmetry of $T_P$: Since invariance of equations of a theory is defined in terms of symmetries of that theory, $T_P$’s equations will, in general, not be invariant with respect to $(\text{Aut}(L)\not\subseteq \Sigma)$. There is simply no invariance condition to be posed on the laws of $T_P$ with respect to $(\text{Aut}(L)\not\subseteq \Sigma)$. Therefore there is, in general, no term $\Xi(x)$ with the transformation property (5.18) in the equations of the theory $T_P$.

Here, the qualifier “in general” refers to the fact that mathematically, the laws of $T_P$ might just happen to be invariant w.r.t. $(\text{Aut}(L)\not\subseteq \Sigma)$ for no apparent reason,
even though the latter does not constitute a symmetry of \(T_P\). Since this is not the case for the known fundamental physical theories, and also incompatible with the adherence of a law of parsimony when formulating physical theories, we exclude this case with the following assumption. Note that this assumption might actually be proven if more details about the physical theory \(T_P\) are assumed.

**Assumption 5.11.** We assume that if \((\text{Aut}(L)^k, \tilde{\sigma})\) is *not* a symmetry of \(T_P\), the spatial derivative terms \(t\) of \(T_P\) are not invariant with respect to \((\text{Aut}(L)^k, \tilde{\sigma})\).

According to Section 5.1, whether \((\text{Aut}(L)^k, \tilde{\sigma})\) is a symmetry of \(T_P\) depends on whether one considers models of consciousness which postulate the physical as closed. If one holds that it only makes sense to consider models of consciousness which do *not* postulate the physical as closed (e.g. because the assumption of the closure of the physical violates a necessary condition for the possibility of a scientific study of consciousness, as argued in Remark [A.1], it follows from Lemma 5.3 or Lemma 5.4 that \((\text{Aut}(L)^k, \tilde{\sigma})\) cannot be a symmetry of \(T_P\). If one furthermore holds the assumptions of Proposition 5.10 to be plausible (which we have tried to argue for in Sections 5.2.1 to 5.2.4), one is forced to accept the existence of (5.14). Using Assumption 5.11 this implies that \(T_P\) does not contain a term (5.14) which transforms as (5.18) for every \(\bar{s} \in \text{Aut}(L)^k\). Therefore, \(\Xi_\mu\) is a *new object* which is not contained in the physical theory \(T_P\). Due to the form (5.14) of this object, and since according to Proposition 5.10 the model’s equations necessarily contain an interaction term between \(\Xi_\mu\) and space-dependent quantities \(\tilde{p}\) of \(T_P\), \(\Xi_\mu\) needs to be interpreted as a new interacting physical field:

**Corollary 5.12.** Any non-trivial model of consciousness \(M\) which

1.) does not postulate the physical closed,
2.) satisfies the technical assumptions of Lemma 5.3 or Lemma 5.4
3.) contains invariant spatial derivatives,
4.) and determines qualia locally

necessarily contains a field (6.14) which is not part of the physical theory \(T_P\) but interacts with the dynamical quantities of the latter.

*Proof.* Proposition 5.10 implies that there exists a field (5.14) which interacts with space-dependent quantities \(\tilde{p}\) of \(P\) and transforms as (5.18) for every \(\bar{s} \in \text{Aut}(L)^k\). Lemma 5.3 or Lemma 5.4 imply that \((\text{Aut}(L)^k, \tilde{\sigma})\) is not a symmetry of \(T_P\). Assumption 5.11 implies that \(T_P\) does not contain a term (5.14) which transforms as (5.18) for every \(\bar{s} \in \text{Aut}(L)^k\).

In summary, Corollary 5.12 shows that any model of consciousness which satisfies a set of plausible assumptions implies that the physical domain is changed in a fundamental way. Whether or not these assumptions are taken to be necessary depends, to a certain extent, on which metaphysical position one appreciates. However, Remark [A.1] and the comments made in Sections 5.2.1 to 5.2.4 make a strong case for these assumptions which is independent of particular metaphysical convictions.

---

This is so because usually, symmetries are deduced from invariance properties of the laws of physical theories.
Consciousness is in the focus of research projects around the globe. Empirical as well as theoretical projects aim to investigate different aspects of experience, ranging from access consciousness or the unity of a conscious scene to phenomenal consciousness or the first-person-perspective [Set07]. The starting point of this paper is the observation that if an aspect of experience is under investigation which cannot be collated over several experiencing subjects (i.e., which cannot be *communicated*), special care is necessary. Any reference to such aspects of experience, be it in a theoretical account or simply by verbal report, is ambiguous and this ambiguity may lead to ill-defined models, conflating empirical predictions and experimental data which identifies the wrong correlate of some aspect of experience.

In order to develop a well-defined scientific methodology which can be applied to all aspects of experience, we have used basic *phenomenological facts* to specify how a formal representation of experience can be constructed. The result is a mathematical space which represents some parts of experience (such as visual experiences or auditory experiences) *completely*, including both the usual objects of investigation in cognitive neuroscience as well as aspects of experience which cannot be collated. This formal representation of experience avoids the usual hard cut between parts of experience which represent a difficulty for the scientific methodology and parts which do not. Both are simply two aspects of the whole of experience, similarly to colour and form being two aspects of a painting, or similarly to position and momentum being two aspects of a quantum state.

We have shown that this mathematical representation of experience allows to quantify the ambiguity involved in any reference to experience precisely. This is sufficient to avoid the problems mentioned above and yields a formal mathematical toolbox which can be applied in empirical or theoretical investigations of consciousness.

In the second part of the paper, we have investigated how individual non-communicable aspects of experience (qualia “sensu stricto” [OAT14]) can be studied scientifically. Since there is a fundamental explanatory gap, i.e. a more genuine “hard problem of consciousness” [Cha95], this question may be considered as equally relevant to the one addressed in the first step.

The main result of the second part of this paper is that formal models of consciousness can address individual non-communicable aspects of experience *if and only if* they carry a specific symmetry group related to the mathematical representation of experience explained above. Because of mathematical details of the action of this symmetry group, models of consciousness can be used to construct empirically well-defined theories of how individual aspects of experience relate to the physical domain *despite* the ambiguity inherent in any reference to the latter. We have found that models of consciousness, defined suitably in terms of the representation of experience constructed in the first step, offer a sufficient formal framework for an empirically meaningful scientific approach to consciousness.

In the third part of this paper, we have generalized the ‘gauge symmetry principle’ of quantum field theory to models of consciousness. This has allowed us to *prove* that the symmetry requirement found in the second part can have substantial consequences for the mathematical structure of the physical theory underlying a model of consciousness.
The results of the first and second part of this paper constitute a grounding of the scientific study of consciousness which is an alternative to other groundings currently in use. It offers a thorough conceptual and mathematical framework in light of which existing models of consciousness can be interpreted and improved, and based on which new models can be constructed.

Acknowledgements: I am grateful for the questions and comments received during presentations of parts of this work at the LPS Colloquium of the Munich Center for Mathematical Philosophy, the Mathematical Institute of the University of Göttingen, the Institute for Theoretical Physics of the University of Hannover, the Modelling Consciousness Workshop in Dorfgastein, the Online Seminar Progress and Visions in the Scientific Study of the Mind-Matter Relation and the Conceptual Foundations of Science Workshop in Tegernsee.

Appendix A. Conceptual Problems of Chalmers’ Grounding

In this appendix, we briefly discuss several conceptual issues of Chalmers’ grounding. These issues are not motivated by metaphysical considerations and are not intended to have metaphysical implications; they simply arise if one wishes to carry out a scientific investigation of consciousness based on Chalmers’ grounding. Problems A.1 and A.2 are most crucial and might make it impossible to apply the grounding.

The abbreviations used in this appendix have been introduced in Section 2. For reasons explained in Appendix A.4, we assume that Assumption (A1) is intended to express the fact that “physical laws already form a closed system” [Cha96, p. 127].

A.1. Closure of the Physical. Much has been written about Assumption (A1) both by David Chalmers himself (e.g. [Cha96, Ch. 5] or [Cha10, Ch. 8 and 9]) and by others (e.g. [Eli09] or [Bis05]). As noted in Section 2, this assumption is crucial for Chalmers’ grounding in order to establish that there are aspects of experience which satisfy (D1).

To date, there is no valid argument which shows that Assumption (A1) is wrong, i.e. that the physical laws of nature cannot “form a closed system”. On the other hand, there also is no valid argument that shows that Assumption (A1) is right, i.e. that the physical laws of nature must form a closed system. This assumption also cannot be backed by analysing opinions or strategies of working physicists, for most physicists are prepared to accept, or even try to find, modifications of the known laws of physics due to yet unknown phenomena (e.g. related to dark matter, to quantum gravity or to dynamical collapse theories, to name just a few). They do not assume that the known physical laws form a closed system. “Physics itself does not imply its own causal closure nor is there any proof within physics of its own completeness, so CoP [causal closure of physics] must be a metaphysical principle” [Bis05, p. 45].

Based on this state of affairs, one might think that both Assumption (A1) as well as its opposite should be compatible with a scientific approach to consciousness. However, this is not the case, as the following remark shows. Despite the fact that the physical laws of nature may form a closed system, it seems that Assumption (A1) is incompatible with a scientific approach to investigate consciousness because it violates a necessary condition for the possibility of the latter.

Note that no reasons are given in either [Cha96] or [Cha10] for why Assumption (A1) should hold true.
Remark A.1. As explained in Section 3.6, the phenomenological grounding allows to construct models of consciousness which postulate the physical as closed just as well as models which do not postulate the physical as closed (several examples of both are given in Section 4). However, it seems that in both Chalmers’ and the phenomenological grounding of the scientific study of consciousness, it does not make sense to assume the closure of the physical because it violates a necessary condition of the possibility of the scientific study of consciousness itself.

The goal of this remark is to explain in detail why this is so. To this end, we use the symbol $Q$ to denote that which is to be studied according to the grounding at hand: In the case of Chalmers’ grounding (CG), $Q$ refers to qualia as defined in (D1) in Section 2 whereas in the case of the phenomenological grounding (PG), $Q$ refers to qualia as defined in Definition 3.8. In both cases, $Q$ thus refers to aspects of conscious experience.

The above claim rests on two premises. First, that a scientific study of consciousness is possible only if scientists can communicate about $Q$ at least to some extent. E.g., they need to be able to agree on $Q$’s definition and existence, need to be able to communicate certain general properties of $Q$ (such as Phenomenological Facts 3.5, 3.11 or 3.13 in the case of PG) or need to be able to record and exchange data related to $Q$. This is the necessary condition for the possibility of the scientific study of consciousness referred to above, which we abbreviate by NC.

The second premise is that communication is always mediated via communication channels $C$ which are elements of the physical domain. To give some examples, consider verbal communication, which is mediated via sound waves, digital communication, which is mediated via electromagnetic signals, or printed texts, where communication is mediated via arrangements of molecules and electromagnetic fields.

Due to the second premise, an assumption concerning the closure of the physical (ACoP) has something to say about communication channels and therefore also about communication itself. If, in the grounding at hand, ACoP is fleshed out in such a way that it restricts the relation between $Q$ and communication channels $C$ to such an extend that communication about $Q$ is impossible, the above claim holds: By the first premise, this implies a violation of a necessary condition of the possibility of a scientific study of consciousness.

Clearly, whether or not this is the case depends on what one takes to constitute ‘communication’ and which conditions one posits as necessary for something to count as ‘communication about $Q$’. To find proper answers to these questions is of course the goal and task of various parts of philosophy, and little of general nature can likely be said without a proper training in the relevant fields. However, by restricting to a very simple situation, we may hope to work with a necessary requirement for ‘communication about $Q$’ to be possible which is acceptable independently of which notion of communication one prefers.

The simple situation which we consider is the prototypical scenario of the mathematical theory of communication. I.e., we consider a situation where one experiencing subject $S_1$ (the ‘sender’) formulates a message $m_1$ which expresses some properties

---

53 This is the original (and arguably more adequate Flo17) name for ‘information theory’ Shad8.
of her experience of \(Q\), such as which particular phenomenal quality she has experienced (in the case of CG) or whether two qualia are similar (in the case of PG). Subsequently, this message is being transferred via a communication channel \(C\) to another experiencing subject \(S_2\) (the ‘receiver’), who after decoding the channel’s signals obtains a message \(m_2\). We abbreviate this scenario by MTCp (‘p’ for ‘prototypical’).

We denote properties of \(S_1\)'s experience of \(Q\) by \(q\) and states of the communication channel \(C\) by \(c\). In what follows, we consider functional dependencies between the quantities \(q, c, m_1\) and \(m_2\). In order to define what constitutes a functional dependency both mathematically and conceptually, we refer to the groundings under consideration:

Both CG and PG’s specification of the task of the scientific study of consciousness includes the formulation of laws or theories concerning the relation of \(Q\) with the physical domain. Given enough further specifications (such as a model of the communication channel or more comprehensive physical laws), these laws or theories should be applicable to the MTCp setup. I.e., we may assume that both PG and CG allow to construct (or even to deduce) mathematical models of the MTCp setup. The details of any such model of course depend on various factors, most importantly on which psychophysical laws (CG) or models of consciousness (PG) one considers. All that matters at this point is that given any such model, we may identify functional relationships between the quantities \(q, c, m_1\) and \(m_2\):

(F1) A quantity \(a \in \{q, c, m_1, m_2\}\) is functionally dependent on a quantity \(b \in \{q, c, m_1, m_2\}\) according to some model of the MTCp setup iff according to this model, \(a\) is a non-constant function of \(b\).

The reasons for focussing on functional dependency in order to argue for the main claim of this remark are threefold. The first reason is that in both CG and PG, ACoP implies a restriction of the functional dependencies which may hold between the quantities \(q, c, m_1\) and \(m_2\). Consider first CG. Here, the various formulations of ACoP differ slightly depending on whether they utilize a notion of causality or not. However, it seems fair to say that they all intend to express the central claim that “physical laws already form a closed system” [Cha96, p. 127]. Together with the second premise introduced above, this implies that in any model of the MTCp setup based on CG, \(c\) cannot functionally depend on \(q\). In PG, ACoP is formulated in Definition 3.31. It implies that any state \(c\) of the communication channel is determined completely by the dynamics of the physical theory \(T_P\), which does not include \(Q\). Therefore, as is the case for CG, in PG ACoP also implies that \(c\) cannot functionally depend on \(q\):

(A2) In both Chalmers’ grounding (CG) and the phenomenological grounding (PG), the assumption of the closure of the physical (ACoP) implies that the states \(c\) of communication channels cannot be functionally dependent on \(q\).

The second reason is that functional dependency also seems to allow to formulate a fundamental necessary condition for ‘communication about \(Q\)’ to be possible:

(C1) A necessary condition for communication between \(S_1\) and \(S_2\) about \(Q\) is that \(m_2\) may depend functionally on \(q\).

\[54\]If this is impossible, i.e. if the assumptions of a grounding are such that an experiencing subject cannot formulate a message which expresses some properties of her experience of \(Q\), this grounding violates the necessary condition NC as claimed. This may be the case for CG, cf. [Cha10, Ch. 9].

\[55\]Here, by ‘constant function’ we simply refer to functions which are formally dependent on \(b\) but whose value remains the same independently of which value \(b\) takes. E.g., \(f(x, y) := x\) is a constant function of \(y\).
The third reason, finally, is that the MTCp is intended to express functional relationships in the first place. In particular, it can be taken to imply by definition that $m_2$ is functionally dependent only on $c$ and acquires additional functional dependencies only via $c$’s functional dependencies.

This concludes the reasoning: A necessary condition of communication about $Q$ in the MTCp setup is that $m_2$ is functionally dependent on $q$. By definition of the MTCp setup it can only be functionally dependent on $q$ via $c$. CG and PG’s ACoP however imply that $c$ cannot be functionally dependent on $q$. Therefore, a necessary condition of communication about $Q$ is violated, which by the first premise above is a violation of a necessary condition for the possibility of a scientific study of consciousness.

Clearly, this reasoning does not yet constitute a formal argument. Several of its suppositions have to be checked carefully for hidden assumptions, which goes beyond the scope of this remark.\[^{56}\] Nevertheless, it is of importance both with respect to Chalmers’ grounding (where it raises a thorough problem) and with respect to the phenomenological grounding (where it is a basis for potential empirical predictions).

We close this remark by pointing out that arguments which try to prove that the closure of the physical cannot hold in light of empirical facts about our experience (most notably written or verbal statements which express some fact about conscious experiences, e.g. bafflement about why consciousness exists [Eli09]) do not seem to be valid. The problem is simply that we may appear to be expressing facts about our conscious experience while in fact we are not. Similarly, we may appear to be communicating about consciousness while in fact we are not. This is the basis for Chalmers’ efforts to develop a theoretical account of how judgements or statements about consciousness can be accounted for despite the closure of the physical, cf. [Cha96, Ch. 5] and [Cha10, Ch. 8 and 9].

In contrast, the claim proposed in this remark simply represents a transcendental argument: Independently of whether reality satisfies the closure of the physical or not, it does not make sense to engage in a scientific study of consciousness if one postulates the physical as closed, because the latter violates a necessary condition of the possibility of the former.

\[\diamondsuit\]

\section*{A.2. Experiments.} An issue also arises with respect to experiments if one postulates that “physical laws already form a closed system” [Cha10, p. 17]: Almost all experiments one might wish to perform are rendered meaningless. The reason is simply that most experimental data (fMRI scans, EEG signals, verbal reports, etc.) is stored on physical devices (hard drives, paper, sound waves etc.) and hence subject to physical laws. If these are postulated to “form a closed system” it follows that the experimental data must be determined by these physical laws alone, independently of which “psychophysical law” [Cha96, p. 127] correctly describes how phenomenal properties depend on physical properties.

To see this in more detail, let us assume that two different psychophysical laws $L$ and $L'$ have been proposed. The idea behind Chalmers’, and in fact any, conception of

\[^{56}\text{To give one example: Above we have borrowed the notion of functional dependency from models of the MTCp setup, which in turn are based on psychophysical laws or models of consciousness. In doing so, we have avoided the difficult question of what a functional dependency actually expresses (i.e. how it is supposed to be defined and interpreted). E.g., when considering $q$, $c$, $m_1$ and $m_2$ as variables, which sort of possible words do they describe? Logically possible worlds, conceivable worlds, some sort of nomologically possible worlds? In what way can the assumptions of a grounding restrict these possible worlds and what effect does this have on functional relationships?}\]
of the *scientific* study of consciousness is that experiments have to be carried out in order to evaluate which of the proposals better describes reality. Accordingly, assume that an experiment has been designed and carried out which purports to answer this question, e.g. by checking predictions based on the laws $L$ and $L'$. Finally, denote by $d$ the dataset produced by this experiment.

The term ‘data’ is applicable to any “putative fact regarding some difference or lack of uniformity within some context” [Flo17], so that one might consider the case where $d$ actually consist of non-physical quantities, e.g. of differences in one’s own experience. However, as soon as the data is stored or processed as usual, e.g. on a hard drive in order to perform statistical analysis, the differences in question have been transformed into “difference or lack of uniformity” of physical quantities. Since almost all experiments, even when dealing with verbal reports or similar indications of conscious experience, perform some sort of statistical analysis, it seems that in almost all experiments, $d$ eventually is a physical data set in this sense. It is ‘stored via’ physical quantities.

If one assumes that “physical laws already form a closed system” [Cha10, p. 17], it follows that all physical quantities, as well the differences or lack of uniformity they exhibit, are determined by the laws of physics alone. Applied to the physical quantities on which $d$ is stored, this statement literally says that the data $d$ is determined by the laws of physics alone. Put differently, due to the fact that the experimental data $d$ is stored on a physical device, closure of the physical implies that the data $d$ is completely independent of whether $L$ or $L'$ or some completely different psychophysical law best describes how experience arises from physical processes.

Thus, in summary, the closure of the physical implies that whatever experiment one performs in order to evaluate psychophysical laws, if it yields data that is stored on physical devices, the result of the experiment is independent of how experience actually arises from physical processes, i.e. independent of that which it seeks to study.

This is conclusion holds true even if we concede that every experiencing subject might interpret the physical dataset $d$ in terms of his/her own experience, so as to give meaning to this set in a way that a philosophical zombie might not, simply because if $d$ is independent of which law $L$ best describes how phenomenal properties arise from physics, the meaning a scientist gives to $d$ will generally be too.

A.3. **Subsumed Notion of Explanation.** Chalmers’ grounding builds on and extends the notion of an explanatory gap that has been introduced by Joseph Levine in [Lev83]. To this end, Chalmers claims that a specific account of explanation covers all notions of explanation that are used throughout natural science: An account in terms of function and structure, cf. [E1] in Section 2. He subsequently shows that there are aspects of experience which do not have any of these two properties, so cannot be explained in terms of natural science as usual. The gist of his grounding is

---

57 If one assumes that communication between two experiencing subjects is mediated via communication channels that are part of the physical domain (cf. Remark A.1), it follows that every scientifically meaningful data needs to be transformed into physical data at some point.

58 One may be able to avoid this last conclusion by insisting that the meaning attributed to $d$ by any experiencing subject is dependent on the law $L$ itself and if one furthermore argues that a conclusion about which law $L$ best fits nature can be deduced from the meaning of $d$, despite $d$ itself being determined independently of the former. At the present stage it seems quite unclear how such an deduction might work, let alone what role an experiment might play in this deduction in the first place.
that they may be addressed by a “new sort of explanation” [Cha96, p. 121] which consists of “new fundamental laws (...) specifying how phenomenal (or protophenomenal) properties depend on physical properties” [Cha96, p. 127].

The question of how scientific explanation is to be defined has occupied many philosophers throughout the 20th century [Woo17]. To find a definition which is general enough to capture the various explanations in science, yet specific enough to exclude scenarios which are clearly not cases of scientific explanation turns out to be a very difficult task. Even basic questions such as whether or not causality is to feature in the definition of explanation (and if yes, which definition of causality), are still largely debated: “There is considerable disagreement among philosophers about whether all explanations in science and in ordinary life are causal and also disagreement about what the distinction (if any) between causal and non-causal explanations consists in.” [Woo17].

This sheds some doubt on Chalmers’ notion of explanation, and the question arises whether (E1) really covers all, or even the most essential, uses of explanation throughout sciences. This is particularly so with respect to physics, whose notion of explanation seems to be a lot more formal than suggested by the terms ‘function’ and ‘structure’ as defined here. E.g., physics does seem to provide notions of explanation which can be applied to general dynamical quantities, whether they describe changes in the behaviour of a system or changes of a more general sort. (Chalmers might even reluctantly agree to this last observation when claiming that “throughout the higher-level sciences, reductive explanation works in just this (E1) way” [Cha10, p. 7], thus, in this quote, avoiding the claim that (E1) also applies to lower-level sciences, such as (presumably) physics.)

This is a problem because the legitimacy of proposing “new fundamental laws” which describe how phenomenal aspects of experience depend on physical properties [Cha96, p. 127], as compared to a reductive explanation in terms of physical accounts, is granted, in Chalmers’ grounding, by the existence of an explanatory gap between phenomenal aspects and contemporary scientific explanation. If scientific explanation is more powerful than Chalmers assumes, the justification of this explanatory gap breaks down and it becomes questionable whether this explanatory gap actually exists. “[A]n explanatory gap (...) cannot be made more precise than the notion of explanation itself” [Lev83, p. 358].

59Recall that the term function refers to “any causal role in the production of behavior that a system might perform” [Cha10, p. 6]. One could interpret this as referring to “any change in the behavior of a system” (cf. Appendix A.4). This could, in turn, be taken to mean “any change in the dynamical properties of a system”, which would change the meaning of the claim that “physical accounts explain only structure and function” [Cha10, p. 105f.] to the following:

“Any account given in purely physical terms will suffer from the same problem. It will ultimately be given in terms of the structural and dynamical properties of physical processes, and no matter how sophisticated such an account is, it will yield only more structure and dynamics. While this is enough to handle most natural phenomena, the problem of consciousness goes beyond any problem about the explanation of structure and function [sic], so a new sort of explanation is needed.” [Cha96, p. 121]

However, most or even all aspects of consciousness are dynamical in nature, which implies that the set of phenomenal aspects of consciousness (Definition (D1)) is, given this redefinition of the term ‘function’, either empty or trivial. Put differently, with this redefinition the grounding implies that all or almost all of conscious experience can be addressed by an “account given in purely physical terms”. What is left out are only non-dynamical aspects of experience (if there are such aspects at all).
A.4. **Causality.** Finally, the question arises of what exactly one should take to constitute causality when applying Chalmers’ grounding. This is so because Assumption [(A1)] as well as Definitions [(E1)] and [(D1)] all relate to causality in an essential way (the latter via the definition of the term ‘function’, cf. Section 2).

This question is widely debated both in physics and in the philosophy of causation \[Sch16\]. It seems fair to say that consensus is missing on basically all aspects of a definition of causality, including basic questions such as which relata a causal relation is to refer to and how, given a choice of relata, causality is defined. Whereas this multitude of possible notions of causality may not matter much if one is concerned with philosophical investigations based on Chalmers’ grounding (one may just restrict to analyses that apply to every notion of causality), it does matter if one wishes to apply the grounding. In particular, if one wishes to model, let alone identify, phenomenal aspects of experience, one does need to know what exactly the Definition [(D1)] amounts to. Since the term ‘function’ used in that definition refers exclusively to causality, the defining property of phenomenal aspects depends on what one takes causality to be.

Connected to questions of how to define causality is the question of the ontological status of causality. Does some definition of causality pertain to “reality” or the universe? (In physicists’ terms: Is causality “fundamental”? Does the universe “obey” one particular definition of causality?) Or is causality rather a tool which can be utilized (by humans, animals, etc. or by information processing systems in general) to describe some parts of reality well to some extent?

Chalmers’ grounding is strongly dependent on which answer one gives to this question. E.g., it determines which sort of “influence” of the phenomenal domain on the physical domain is compatible with the definitions of the grounding, or which type of condition the definition of phenomenal aspects of consciousness constitutes. One may ignore this problem as long as one applies the grounding to theories of the physical domain which incorporate some notion of causality, such as, arguably, abstract neural networks. However, if one wishes to apply the grounding to fundamental physical theories, whose laws do not refer to, or come equipped with, any notion of causality, this question cannot be ignored.

These issues can be avoided completely if one takes the various uses of the term “causality” in Chalmers’ grounding to jointly mean that the physical domain is not changed in any way by phenomenal aspects of consciousness, i.e., that the various uses of causality simply amount to ensuring that the “physical laws already form a closed system” \[Cha96, p. 127\]. This seems to be the actual intention of the author in \[Cha96\] and \[Cha10\], which is why we have fixed this interpretation in the beginning of this appendix.

---

60E.g., \[Pea09\] holds that “[i]f you wish to include the entire universe in the model, causality disappears because interventions disappear – the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece in – namely, the focus of investigation. The rest of the universe is then considered out or background and is summarized by what we call boundary conditions. This choice of ins and outs creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about ‘outside intervention’ and hence about causality and cause-effect directionality.” \[Pea09, p. 419f\] “What we conclude (...) is that physicists talk, write, and think one way and formulate physics in another.” \[Pea09, p. 407\]
REFERENCES

[Bis05] Robert C. Bishop. The hidden premise in the causal argument for physicalism. *Analysis*, 66(1):44–52, 2005.

[Cha95] David Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 1995.

[Cha96] David Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford Univ. Press, New York, 1996.

[Cha10] David Chalmers. *The Character of Consciousness*. Philosophy of Mind. Oxford University Press, New York and Oxford, 2010.

[Cra02] Carl F. Craver. Structures of scientific theories. In Peter Machamer and Michael Silberstein, editors, *The Blackwell Guide to the Philosophy of Science*, chapter 4, pages 55–79. Blackwell Publishers, 2002.

[DCN11] Stanislas Dehaene, Jean-Pierre Changeux, and Lionel Naccache. *The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications*, pages 55–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Den88] Daniel C. Dennett. Quining qualia. In A. Marcel and E. Bisioch, editors, *Consciousness in Contemporary Science*. Oxford University Press, 1988.

[Den93] Daniel C. Dennett. Quining qualia. In Alvin I. Goldman, editor, *Readings in Philosophy and Cognitive Science*. MIT Press, 1993. First printed in [Den88].

[DKC98] Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24):14529–14534, 1998.

[DN01] Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1):1 – 37, 2001. The Cognitive Neuroscience of Consciousness.

[EB98] Editors of Encyclopædia Britannica. Space. In *Encyclopædia Britannica*. Encyclopædia Britannica, inc., 1998.

[EFT84] H.D. Ebbinghaus, J. Flum, and W. Thomas. *Mathematical Logic*. Undergraduate Texts in Mathematics. Springer New York, 1984.

[Eli09] Avshalom C. Elitzur. Consciousness makes a difference: A reluctant dualist’s confession. In A. Batthyany and A. C. Elitzur, editors, *Irreducibly Conscious: Selected Papers on Consciousness*. 2009.

[Flo17] Luciano Floridi. Semantic conceptions of information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[Giul09] Domenico Giulini. Concepts of symmetry in the work of Wolfgang Pauli. In Harald Atmanspacher and Hans Primas, editors, *Recasting Reality: Wolfgang Pauli’s Philosophical Ideas and Contemporary Science*, pages 33–82. Springer Berlin Heidelberg, 2009.

[Gri18] Peter Grindrod. On human consciousness: A mathematical perspective. *Network Neuroscience*, 2(1):23–40, 2018.

[HP14] Donald D. Hoffman and Chetan Prakash. Objects of consciousness. *Frontiers in Psychology*, 5:577, 2014.

[Hus60] Edmund Husserl. *Cartesian Meditations*. Martinus Nijhoff Publishers, 1960.

[Jos15] J. Jost. *Mathematical Concepts*. Springer International Publishing, 2015.

[KR15] Kobi Kremnizer and André Ranchin. Integrated information-induced quantum collapse. *Foundations of Physics*, 45(8):889–899, Aug 2015.

[Kue10] R. Kuehni. Color spaces. *Scholarpedia*, 5(3):9606, 2010.

[Lee02] John M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2002.

[Lev83] Joseph Levine. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(4):354–361, 1983.

[Mar03] Christopher A. Martin. On continuous symmetries and the foundations of modern physics. In Katherine Brading and Elena Castellani, editors, *Symmetries in Physics: Philosophical Reflections*, page 29–60. Cambridge University Press, 2003.

[Mas16] Jonathan W. D. Mason. Quasi-conscious multivariate systems. *Complexity*, 21(S1):125–147, 2016.
