Topics in Mathematical Consciousness Science

Johannes Kleiner

Graduate School of Systemic Neurosciences

LMU Munich

Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

18 December 2024

Supervisor
Prof. Dr. Stephan Hartmann
Chair of Philosophy of Science
Munich Center for Mathematical Philosophy

First Reviewer: Prof. Dr. Stephan Hartmann
Second Reviewer: Prof. Dr. Ophelia Deroy

Date of Submission: 18 December 2024
Date of Defense: 30 May 2025

Brief Summary

The scientific study of consciousness, also referred to as consciousness science, is a young scientific field devoted to understanding how conscious experiences and the brain relate. It comprises a host of theories, experiments, and analyses that aim to investigate the problem of consciousness empirically, theoretically, and conceptually. This thesis addresses some of the questions that arise in these investigations from a formal and mathematical perspective. These questions concern theories of consciousness, experimental paradigms, methodology, and artificial consciousness.

Regarding theories of consciousness, the thesis contributes to the understanding of the mathematical structure that some of the formal theories in the field propose. The work presented here targets the theory of consciousness known as Integrated Information Theory (IIT) and the neuroscientific theory known as Predictive Processing Theory or Free Energy Principle in its Active Inference form (AI-PP). The thesis provides axiomatic definitions of the mathematical structures that constitute these theories, and uses these definitions to address some of the open questions surrounding the theories. For AI-PP, this includes a rigorous derivation of the formula for Active Inference via Free Energy minimisation and a proof of compositionality of Free Energy. For IIT, this includes resolutions of some of the criticisms of IIT's formal scope and applications, but also the identification of new issues that concern the formalism and its derivation. When possible, the definitions are provided in the mathematical framework of category theory.

Regarding experiments, the thesis addresses the main paradigm for testing and falsifying theories of consciousness currently applied in the field. This paradigm consists of comparing the conscious experience that a theory predicts with the conscious experience that is inferred from behavioural data or report by use of measures of consciousness. The thesis provides a formal model of this paradigm and shows that under a certain condition—if inference and prediction are independent—, any minimally informative theory of consciousness can always be falsified. This is deeply problematic since the field's reliance on report or behaviour to infer conscious experiences, in conjunction with the general structure of most contemporary theories of consciousness, implies such independence. This observation provides the exact formal underpinning of the well-known unfolding argument. The thesis analyses the origin of the problem and identifies precisely which changes are required to avoid this problem in future research. The thesis furthermore shows that the problem of falsifying theories of consciousness, and of empirical comparisons of theories of consciousness more generally, follows from a pervasive closure paradigm in consciousness science, that consists of taking a neuroscientific account of the brain as input to a theory of consciousness, so as to explain what

consciousness is, without allowing for modifications or adaptations of the neuroscientific account that would accommodate consciousness as part of the brain's functioning. As is shown in the thesis, this paradigm has implications that point to a fundamental need of revision.

Regarding methodological and conceptual questions, the thesis contributes to the foundations of structural research in consciousness science. Structural research aims to use mathematical structures or mathematical spaces, instead of verbal descriptions or simple categorisations, to represent conscious experiences scientifically, for example when building theories of consciousness, or when exploring new empirical avenues to measure consciousness. Despite considerable advances in this realm, there was, prior to this thesis, no explicit definition of what a mathematical structure of conscious experience should be; that is, how the attribution of mathematical structure to conscious experiences should be systematically understood. Perhaps the most important contribution of this thesis to the field is to propose such a definition. The definition, a structural concept, extends existing approaches wherever available, and provides a basis for developing a common formal language to study consciousness, bridging developments as far apart as psychophysics and phenomenology. In addition, and independently of this proposal, the thesis offers a critical analysis of which metaphysical premises need to be presumed in structural research, whether the use of particular formal tools (such as structure-preserving mappings or homomorphisms) is justified, and how structural theories of consciousness could otherwise be built in the first place. An attempt to expand the results from consciousness to more general problems in philosophy of science is made in the context of the well-known Newman problem.

Regarding the question of artificial consciousness—can AI feel?—, the thesis contributes two results that take the form of no-go theorems, as known from physics. The first no-go theorem shows that if consciousness is relevant for the temporal evolution of a system's states—if it is dynamically relevant—, then contemporary AI systems cannot be conscious. That is because AI systems run on CPUs, GPUs, TPUs or other processors which have been designed and verified to adhere to computational dynamics that systematically preclude or suppress deviations. The second no-go theorem is situated in the context of computational functionalism, a view which posits that consciousness is a computation. The theorem shows that if computational functionalism holds true, consciousness cannot be a Turing computation. Rather, it must be a novel type of computation that has recently been proposed by Geoffrey Hinton, called mortal computation.

This thesis is part of a global effort to pioneer a mathematical perspective in consciousness science, now called *Mathematical Consciousness Science*. The hope behind the research carried out in this PhD is to illustrate the power and usefulness of mathematical approaches in different areas of consciousness science, and in doing so, to lay the foundations for future mathematical work that complements and supports empirical and theoretical work in the further development of this exciting field.

Contents

1. Introduction

Consciousness is the last major frontier of known science. Its scientific investigation, known as *Scientific Study of Consciousness*, or simply *Consciousness Science*, is concerned with the question of how conscious experience—often described as "what it is like" for an organism or system to be in some state (Nagel, 1974)—relates to the brain, or the subject matter of the sciences more generally.

It is perhaps somewhat surprising, from an external point of view, that consciousness science comprises a host of mathematical models, methods, and questions. This is the case, on the one hand, because consciousness science builds on mathematical models and mathematical methods developed in other sciences, for example models of the brain, analysis techniques, modelling procedures, or statistical tests. And it is the case, on the other hand, because consciousness itself is amenable to mathematical description and mathematical representation, for example in terms of phenomenal spaces (cf. Section 1.3). As a result, the study and exploration of mathematical topics has become a notable task in consciousness science.

Mathematical reflection upon the methods and tools used to study consciousness was part of consciousness science from the start. One goal of this PhD was to build upon this pioneering work to help establish the scattered exploration of these topics within consciousness science, as well as the enormous interest and contributions to this topic from formal disciplines like physics, mathematics, computer science, and mathematical philosophy, as a full-fledged area of study.

Thanks to the help of and collaboration with a large number of the most outstanding researchers, both in philosophy and the sciences, much progress towards this goal has now been achieved. The area of study within consciousness science that is devoted to the application and study of formal and mathematical methods is now known as *Mathematical Consciousness Science*.[1] It features a dedicated international conference series, various special issues, and seminars and workshops throughout the globe. More importantly, though, it features a community of 200+ researchers whose research and activities will likely play a significant role in humankind's scientific understanding of consciousness.

Mathematical consciousness science is to consciousness science what mathematical physics is to physics, what mathematical biology is to biology, and what mathematical neuroscience is to neuroscience.[2] It is the application and study of formal and math-

[1] The term was first introduced as the title of an online seminar series '*Mathematical Consciousness Science – An online seminar series exploring the role of mathematics in the scientific study of consciousness*,' to which about 1300 researchers had subscribed. In conjunction with other events and endeavours, the series led to the foundation of the *Association for Mathematical Consciousness Science*.

[2] An equally valid metaphor would be to say that mathematical consciousness science is to consciousness

ematical methods as applied in, or relevant to, the scientific study of consciousness. Because mathematical methods and mathematical questions appear in experimental, theoretical, conceptual, and methodological domains in consciousness science, mathematical consciousness science comprises experimental, theoretical, conceptual, and methodological questions.

The research carried out as part of this PhD makes contributions to all of these domains, in vastly different ways. The contributions range from case studies that illustrate how mathematical methods can be used in a specific area (as for example the case for modelling of experiments on consciousness, cf. Section 1.2), to foundational work that addresses a comprehensive body of literature in consciousness science (as for example the case for the analysis of mathematical spaces of conscious experiences, cf. Section 1.3).

The goal of this chapter is to introduce and review these contributions, and thereby to provide the reader with one initial perspective of what mathematical consciousness science is, and how it can contribute to the scientific study of consciousness at large. It is important to note, however, that the chapter does not provide a comprehensive review or conclusive assessment of mathematical consciousness science. It only mentions that part of the literature on the topic that is relevant to the questions pursued within this PhD.

If anything, the term *Mathematical Consciousness Science* should be taken to refer to the questions which members of the community of researchers that work on this topic ask and answer. Much like it was impossible to foresee the developments of modern mathematical physics prior to the 20^{th} century, it is impossible to foresee what mathematical consciousness science might grow to become. The following introduction to some of the contemporary themes of mathematical consciousness science should therefore be understood as outlining a very preliminary picture.

This PhD is a cumulative thesis, meaning that it consists of individual papers, which are presented in Chapters 2 to 11. The papers fall into four broad classes of work in mathematical consciousness science: research on theories of consciousness, on experiments, on methodology, and on the question of artificial consciousness. The present chapter constitutes an introduction and review of the individual projects carried out within this PhD, and with that an introduction to these four branches of mathematical consciousness science. But every paper can also be understood without the introduction provided in this chapter; in case of interest, the reader can jump right into the corresponding chapter. An evaluation of the progress that has been achieved within this PhD, and outlook onto the future, is presented in the conclusion of this thesis in Chapter 12.

science what theoretical physics is to physics. Theoretical physics and mathematical physics largely study the same subject matter but differ in the importance that is given to mathematical properties of mathematical objects. For example, theoretical physicists do not normally worry about the existence of derivatives of mathematical functions, whereas mathematical physicists do.

1.1. Research on Theories of Consciousness

Theories of consciousness, also called models of consciousness, are hypotheses about how conscious experiences and the subject matter of the sciences, most notably the brain, relate. They are usually required to be substantive and non-trivial, and are either derived from experimental data or meaningful conceptual assumptions. In this section, we discuss what mathematical consciousness science contributes to the research on theories of consciousness.

1.1.1. Improving & Clarifying Theories

A first major task which mathematical consciousness science has taken up in regard to theories of consciousness is to improve and clarify them.

Integrated Information Theory Consider, as an example, *Integrated Information Theory* (IIT) (Oizumi, Albantakis, & Tononi, 2014). IIT is considered one of the leading models of consciousness and aims to describe both the quality and quantity of the conscious experience of a system, such as the brain, in a particular state.

IIT comprises two main parts. On the one hand, a conceptual part that spells out IIT's metaphysical presumptions, as well as a set of assumptions that are taken to characterise the essential properties of conscious experiences in full. The latter are referred to as IIT's 'axioms'. On the other hand, it comprises a complex and rather complicated set of mathematical equations that determine the conscious experience of any system, such as the brain, based on a formal description of the system. This formal part of IIT constitutes the actual hypothesis about how the subject matter of the natural sciences (e.g., the brain) and conscious experiences relate. The conceptual part essentially serves as a justification of the formal part: the formal part is meant to be derived from the axioms and metaphysical premises of the conceptual part.

What mathematical consciousness science can contribute to the development and public understanding of IIT is:

- **(a)** the explication and clarification of which mathematical object it is that the equations and formal concepts of IIT actually describe,
- **(b)** the exploration and assessment of problems of IIT's formal constructions, in particular based on the clarification of the mathematical structure uncovered in **(a)**, and
- **(c)** ways to define the formal content of IIT in terms of more appropriate mathematics, both to propose improvements of the theory and to make it easier to understand.

Task **(a)** has been carried out as part of this PhD by Kleiner and Tull (2021) and is presented in **Chapter 2**. The goal of this work was to uncover the mathematical object that underlies the formal descriptions and equations of IIT 3.x, meaning: of any published paper (including supplementary material) that has been published by the lab that

develops IIT after the IIT 3.0 paper (Oizumi et al., 2014) and before IIT 4.0 was first proposed in parts in (A. Haun & Tononi, 2019).

The result of this work is a detailed description and definition of the mathematical mapping that constitutes the formal part of IIT. This mapping maps every formal description of a system, together with a state thereof, to a space of conscious experiences, and element thereof (Figure 2.1).

A surprising discovery in this respect was that much of the mathematical structure that appears to be essential for IIT's description of conscious experience in terms of formal spaces is actually auxiliary and merely derives from the particular notion of (network-like) classical systems that has been applied in previous expositions of the theory. The mathematical investigation carried out as part of task **(a)** allows to delineate between the essential and auxiliary structure. This matters, for example, for assessments of IIT's phenomenological implications, as well as for theoretical work that attempts to put Global Neuronal Workspace Theory (GNWT) (Dehaene, Changeux, & Naccache, 2011) on a par with IIT as far as explanatory scope is concerned.

Task **(b)**—the exploration of reasons to criticise IIT's formalism—has been a prominent and important part of the literature on IIT since it's full formalism was first proposed by Oizumi et al. (2014), cf. for example (A. B. Barrett & Mediano, 2019; A. B. Barrett, 2014; Moon & Pae, 2018; Cerullo, 2015). As part of this PhD, Kleiner and Hoel (2021) have considered a particularly prominent criticism, known as 'unfolding argument' (Doerig, Schurger, Hess, & Herzog, 2019), as well as consecutive investigations of IIT's scope (Michel & Lau, 2020), methodology (Negro, 2020), mathematical framing (Tsuchiya, Andrillon, & Haun, 2020) and testability (Kleiner, 2020a; Ganesh, 2020; Hanson & Walker, 2021).

The result of Kleiner and Hoel (2021)'s investigation, presented in **Chapter 5**, shows that there is a fundamental issue with testing IIT that derives both from IIT's mathematical formalism and the typical paradigm of testing theories of consciousness.

Importantly, this is not an issue that pertains to IIT alone. Rather, this issue appears for all major theories of consciousness presently proposed, in a nutshell because much like one can, in theory, substitute any recurrent system (which is conscious according to IIT) by a feed-forward system (which isn't conscious according to IIT), while keeping the input-output mapping of the system the same, one can in theory substitute any part of a system (for example, a global workspace) by a look-up-table device without changing the input-output mapping of the system as a hole. Substituting a recurrent network by an "unfolded" feed-forward network as in (Doerig et al., 2019), or a finite automaton by an isomorphic finite automaton as in (Hanson & Walker, 2019) are only special cases of a huge class of substitutions that can be performed. Therefore, there is a fundamental issue with falsification of *any* theory of consciousness, if theories of consciousness are understood as presently conceived.

We outline the implications of the problem discovered in (Kleiner & Hoel, 2021) in more detail in Section 1.2, and discuss which steps are necessary to resolve this problem in Section 12.1 of the conclusion. Readers with an interest in this issue are directed to Chapter 5.

First steps towards Task **(c)** have been carried out by Tull and Kleiner (2021) and are presented in **Chapter 3**. The goal of this project was to consider IIT in the context of the powerful and elegant mathematical language of *category theory*. To this end, we have demonstrated how integrated information and other key notions from IIT can be studied within the simple graphical language of process theories (symmetric monoidal categories). As in the work on the mathematical structure of IIT, our desideratum was to stay true to the definitions of IIT 3.x provided in the literature.

The result of this work allows IIT 3.x to be generalised to a broad range of physical theories and sets the foundation for a categorical definition of IIT. A full categorical version of IIT that presents the theory in terms of a functor, however, requires breaking with the formalism of IIT that is published in the literature, and hence is not available to date. The exploration of IIT's relation to category theory, however, is thriving, see for example (Tsuchiya & Saigo, 2021; Tsuchiya, Phillips, & Saigo, 2022; Tsuchiya, Saigo, & Phillips, 2023) and (Prentner, 2024a).

There are many more questions for mathematical consciousness science to consider in relation to IIT, and it is likely that ultimately, IIT can only overcome the various criticisms of its formal structure that have been proposed if it engages with the contributions that are made as part of mathematical consciousness science, most notably those that suggest improvements of the theory. For example, as part of the investigation of the mathematical structure of IIT, Kleiner and Tull (2021) have made a proposal of how IIT's formalism could be amended to overcome the criticism put forward in (A. B. Barrett & Mediano, 2019), see Section 2.11. The amended definition of IIT proposed in this section furthermore is such that the problem discussed in (Hanson & Walker, 2023) cannot occur, qua definition. Both of these proposals could be incorporated into the further development of IIT as part of an effort to respond to and resolve worries that exist in the community of researchers that engage with IIT.

Predictive Processing and Active Inference A second example of improvement and clarification work concerns Predictive Processing Theory (PP) and its Active Inference doctrine, also known as 'Free Energy Principle' (Parr, Pezzulo, & Friston, 2022). While not itself a theory of consciousness, this is arguably a first comprehensive theory of brain function. Because PP and Active Inference aim to offer one coherent principle that explains phenomena as diverse as perception, cognition, planning and action, a connection to conscious experience is not surprising.

While there are comparably simple conceptual ideas that afford a substantial understanding of the theory—prediction, prediction error, prediction error estimation, precision, and so fourth—, the theory is in fact a formal theory of the brain, and only a formal account can grasp the theory in full (Buckley, Kim, McGregor, & Seth, 2017; Parr et al., 2022). What is more, recent expositions of the theory have moved away from formal structures where concepts like prediction error still play an important role, and towards a formal structure that is independent of, and more general than, these ideas, most notably the 'Factor Graph' formulations (De Vries & Friston, 2017). A mathematical exposi-

tion and analysis are therefore helpful not only for inner-theoretic purposes, but also to make the theory accessible for further theorising, in particular where consciousness is concerned.

In **Chapter 4**, Tull, Kleiner, and Smithe (2023) provide a categorical formulation of Predictive Processing Theory (PP) with Active Inference, expressed in terms of a simple diagrammatic formal language known as string diagrams that define a monoidal category from the branch of mathematics known as category theory. This research includes diagrammatic accounts of generative models, Bayesian updating, perception, planning, Active Inference, and Free Energy, as well as a diagrammatic derivation of the formula for Active Inference via Free Energy minimisation. As part of this project, we also established compositionality of Free Energy, allowing Free Energy minimisation to be applied at all levels of an agent's generative model. Aside from aiming to provide a helpful graphical language for those familiar with Active Inference, the goal was also to provide a concise formulation and introduction to the Active Inference framework for use in mathematical consciousness science.

The hope behind the research carried out in Chapter 4 is to provide a mathematical basis that allows to formulate hypotheses about how PP and Active Inference relate to conscious experience in concise and rigorous terms. This is relevant to understanding and clarifying the various hypotheses (cf. for example (Miller, Clark, & Schlicht, 2022)) and methodological ideas (cf. for example (Seth & Hohwy, 2021)) that have been put forward in this context, and constitutes a foundation for future mathematical research on computational phenomenology, cf. Section 12.3.

The Important Case of Non-Formal Theories Integrated Information Theory and Predictive Processing are, at the present stage of development, of particular interest to mathematical consciousness science because they are the only formal theories within the Overton window of consciousness science at large.[3] But it is important to note that the task of improving and clarifying theories of consciousness also concerns theories which are not presented in mathematical form at the present time.

One reason for an interest in and possible contribution to non-formal theories is that many theories of consciousness employ what could be called 'proto-formal' concepts: concepts that allude or refer to formal notions, but are not presented in a formal form. Another reason is that detailed descriptions of neuronal dynamics and brain functions are formal in nature, and if a theory of consciousness claims that consciousness supervenes on, or is identical to, neuronal dynamics or a brain function, it must take their formal structure into account. Either way, formal ideas, concepts and definition are already part of the theories, albeit mostly not in an explicit way.

Consider, as an example, Global Neuronal Workspace Theory (GNWT), which posits that a system has conscious representations only if two necessary conditions are satisfied. The first necessary condition is that the system has "two main computational spaces, each characterized by a distinct pattern of connectivity" (Dehaene et al., 2011).

[3] Several other formal theories of consciousness exist, for example (L. Blum & Blum, 2022) or (Mason, 2021), to mention two.

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 other 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" (ibid.), cf. (Kleiner, 2020b) for a more detailed summary and first formal exposition. The second necessary condition is 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" (Dehaene et al., 2011).

This characterisation of the theory is good enough for contemporary purposes and contemporary experimental investigations. But for the theory to properly handle the question of consciousness in organisms and systems that differ from the standard case of healthy humans, the theory must specify which structure, precisely, counts as a computational space of each kind, and what the necessary "patterns of connectivity" are. Computational spaces and patterns of connectivity are formal concepts, hence ultimately a formal specification is in order.

The clarification, improvement or construction of the mathematical structure of existing theories of consciousness is particularly important in the context of Artificial Intelligence (AI), when investigating the possibility of AI consciousness. Because AI systems are formal systems, a rigorous application of theories of consciousness to AI systems cannot do without such formal expositions.

1.1.2. Building New Theories

In the previous section we have reviewed ways in which mathematical consciousness science can (and to some extent already has) contributed to the study of existing theories of consciousness. A task of equal importance for mathematical consciousness science is work on proposing new theories or models of consciousness. New proposals may be in need of mathematical approaches because:

- **(a)** they address brain functions or neural dynamics on a level of precision that is not amenable to non-formal descriptions,
- **(b)** they address the subject matter of the sciences on scales other than the brain, or
- **(c)** because they rely on principles which cannot be precisely expressed in terms of natural or near-natural language.

While several of the projects carried out within this PhD were concerned with the question of how to build theories of consciousness (cf. Section 12.1 in the conclusion), no new models or theories of consciousness have been proposed as part of this PhD.

1.2. Research on Modelling Experiments

Consciousness science is an inherently empirical discipline. Its progress rests on empirical observation in carefully designed experiments that live up to the highest statistical and methodological standards. While mathematical consciousness science is not concerned with running experiments, it can contribute to the task of designing and abstractly analysing experiments.

1.2.1. Measures of Consciousness & C-Tests

Because consciousness is not publicly observable "just like that", running an experiment that targets conscious experiences differs substantially from experiments in other sciences: it requires means to infer information about the conscious experience of a subject in experimental trials. Such means are called *measures of consciousness*. Simply put, measures of consciousness are "consciousness detection procedures" (Michel, 2023) that can be used to determine whether a subject in an experimental trial has experienced a stimulus consciously or not. There are various measures designed for different paradigms. A simple and effective measure is subjective report: asking a subject whether it has experienced a stimulus consciously or not. But many other measures have been developed as well, in particular to target close-to-threshold stimulus conditions, where subjective reports become unreliable.

A closely related concept has recently been dubbed 'C-tests' (Bayne et al., 2024). C-tests are means to infer whether a system is conscious at all, meaning: whether it has conscious experiences at all, or not. Like measures of consciousness, C-tests make use of empirical experimental data obtained from individual systems and organisms, but unlike measures, they do not seek to infer information about the particular conscious experience in experimental trials. Rather, they aim to test whether a system has conscious experiences at all. Being able to test whether a system is conscious has a huge clinical importance, and matters largely for ethical, judicial and governance questions.

Measures of consciousness and C-Tests are essential for consciousness science to make progress. That is the case because given the contemporary paradigm of constructing neuroscientific theories of consciousness (more on that paradigm in Section 12.1 of the conclusion), they offer the only means to test neuroscientific theories of consciousness in a lab.[4]

Prima facie, neither measures nor C-Tests are mathematical in nature. They are based on subjective reports, objective performance or behavioural measures. What mathematical consciousness science can contribute to the study of measures and C-Tests, however, is to model them formally in conjunction with other hypotheses, for example concerning theories of consciousness.

[4] Contemporary neuroscientific theories of consciousness predict the conscious experience that a system has in a particular state. They do not, or are not usually taken to, posit changes in brain dynamics or brain functions. Hence they can only be tested by comparing the prediction the theory makes with the state of consciousness inferred from a measure of consciousness or C-Test.

Such an analysis is provided in (Kleiner & Hoel, 2021), and presented in **Chapter 5**. Because both measures of consciousness and C-tests operate on data obtained in experimental trials, they can both be modelled in terms of a formal mapping

$$\inf : \mathcal{O} \to E \,,$$

where $\mathcal{O}$ denotes a set of experimental data sets, and E denotes an abstract notion of states of consciousness, which depending on the application can be either states that target individual experiences, or states that are meant to assess whether as system has consciousness at all. In the case of measures of consciousness, $\mathcal{O}$ comprises data obtained in individual experimental trials, for example in contrastive analysis designs; in the case of C-Tests, $\mathcal{O}$ denotes data obtained in testing a subject of interest, for example behavioural indicators of a subject in the case of a non-responsive wakefulness test; the mapping $\inf$ then represents the particular rules and operations that result in 'conscious' vs. 'unconscious' judgements.

What we found in (Kleiner & Hoel, 2021) was that if such consciousness inference procedures are independent from a theory's prediction—a situation which prima facie one would think is ideal—a very counter-intuitive result follows: for every correct inference of an experience, one can modify the part of the system that matters for consciousness so as to obtain a different, non-overlapping prediction while keeping the inferred state constant. Hence the theory must be false, or (if no correct inference exists) untestable. This analysis provides the exact formal underpinning of the unfolding argument that criticises IIT (Doerig et al., 2019). The analysis shows that the argument applies to experimental paradigms, which comprise both theory and measurement, rather than theories alone, and applies to a wide range of theories. More details of this problem, and ways to resolve it, are presented in the conclusion in Section 12.1.2.

1.2.2. The Closure Paradigm

A second example of modelling experiments in consciousness science is (Kleiner & Hartmann, 2023), presented in **Chapter 6**. Unlike the research reviewed in the previous section, this work does not model the inference process via measures of consciousness or C-tests. Rather, it models experiments on a more fundamental level, the level of data collection and data storage.

What this work shows is that a central paradigm that spans experimental and theoretical work in consciousness science needs revision: the paradigm that consciousness science is to take a neuroscientific account of the brain as "input", so as to explain what consciousness is, without amending or adding to this input—without amending or adding to the wealth of neuroscientific knowledge, that is. This 'closure paradigm', as one might call it, is at the heart of both identity theories and functionalist theories of consciousness, and is intimately related to discussions of physicalism and the closure of the physical in philosophy of mind.

Chapter 6 shows that the closure paradigm conflicts with the testability of theories of consciousness. The underlying intuition is simple: if theories don't amend, or add

to, the neuroscientific model of the brain, they cannot account for how experimental data, which relies on reports or behavioural indicators, all of which are subject to the neuroscientific model, speaks in favour of one, rather than another theory.

This has profound implications for how theories of consciousness should be formulated, cf. Section 12.1. At the very least, consciousness needs to be *dynamically relevant* with respect to a reference neuroscientific model, meaning: it must be relevant to the dynamical evolution of a neuroscientific system over and above the dynamical evolution prescribed by the reference model, for otherwise two theories of consciousness cannot disagree about the report that should ensue in experimental trials.

Put in somewhat general terms, this result could be read as a requirement for consciousness to have genuine causal powers,[5] but it is important to note that these need not be extra-physical. On the contrary, it is specifically in the case of physicalist and neuroscientific assumptions that the full force of this result applies: consciousness needs to be understood as a full-fledged physical part or process of the brain; it cannot be subjugated to an epiphenomenon of sorts.

1.3. Research on Conceptual and Methodological Questions

Concepts and methods are essential for any science to move forward. As Daniel Dennett once put it, "there is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination" (Dennett, 1995, p. 21). This is particularly true of consciousness science, where a large number of concepts have been developed in order to refer to the target phenomenon and make it accessible to scientific analysis. This work is all but finished, and research on new concepts to describe, refer, or represent (parts of) the target phenomenon in consciousness science is a particularly important contribution of philosophy of mind to the science of consciousness.

Good concepts are required to develop rigorous experiments and theories, and are essential to avoid mistakes in theorising, cf. e.g. (Nida-Rümelin, 2018). Conceptual work is an essential but often overlooked ingredient in pushing the boundaries of scientific knowledge. Correspondingly, a third pillar of how mathematical consciousness science can contribute to the scientific study of consciousness is the exploration and analysis of formal concepts.

A particularly noteworthy development related to formal concepts in consciousness science is the introduction of mathematical spaces and mathematical structures in order to describe or represent conscious experiences as part of the scientific methodology. While pioneering work in this respect has been carried out right in the initial phases of the field (A. Clark, 1993; Rosenthal, 2010), recent years have seen applications of spaces and structures in virtually every subdiscipline that is part of consciousness sci-

[5] There are many different interpretations of what causal language should mean, cf. for example (Beebee, Hitchcock, Menzies, & Menzies, 2009). This is why the concept of dynamical relevance is formulated without reference to causation.

ence. There are clear signs of a *structural turn* (Kleiner, 2024) in consciousness science that might change the field fundamentally.

A major part of this PhD was devoted to analysing and exploring the application of structures and spaces to conscious experience. The central question in these developments, spanning Chapters 7 to 9, was what claims about the mathematical structure of consciousness *should be taken to mean*. This is the question of which conditions or definitions should or need to be subsumed so as to allow for a meaningful application of mathematical concepts in proposals as diverse as quality spaces (A. Clark, 1993; Rosenthal, 2015; Lee, 2021), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021), qualia structure (Kawakita, Zeleznikow-Johnston, Tsuchiya, & Oizumi, 2023; Kawakita, Zeleznikow-Johnston, Takeda, Tsuchiya, & Oizumi, 2023; Tsuchiya et al., 2022), Q-spaces (Chalmers & McQueen, 2022; Lyre, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink, Kob, & Lyre, 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). We refer to these proposals jointly as proposals of *mathematical structures of conscious experience*.

Chapter 7, published as (Kleiner, 2024), provides an analysis of three popular contemporary ideas in consciousness science that might have the potential to strongly shape initial developments in a structural turn, but which are in fact misunderstandings or wrong. These ideas concern **(a)** the conflation of structural and structuralist agendas, **(b)** unjustified assumptions in using isomorphisms to understand or model the relation between neural substrate and conscious experience, and **(c)** conflation of mathematical structure that originates from laboratory operations or mathematical convenience with structure that actually pertains to conscious experience.

Chapter 8, published as (Kleiner & Ludwig, 2024), provides an analysis of existing definitions of mathematical structures of conscious experiences, most notably of those in the context of quality spaces. It identifies problems in existing approaches and offers a new proposal, built on the shoulder of existing proposals, of how to define structures of conscious experiences, such as phenomenal spaces, quality spaces, qualia spaces, and other of the above-mentioned concepts. The work presented in this chapter aims to provide an improved foundation for structural research in consciousness science to move forward, outlined in Sections 12.2 and 12.3 in the conclusion.

Chapter 9, finally relates the new proposal developed in Chapter 8 to research on structural methodologies in philosophy of science, most notably the so-called Newman Problem (Newman, 1928; Frigg & Votsis, 2011). The chapter shows that the proposal developed in Chapter 8 is a solution of the Newman problem that has a number of advantages over existing approaches. In a sense, this chapter identifies the full force of the developments in Chapter 8.

The hopes and visions behind this research, and how it might help improve consciousness science in the years to come, are described in the conclusion, Chapter 12.

1.4. Research on Artificial Consciousness

In light of the vast developments of Artificial Intelligence (AI) in recent years, questions pertaining to a mind of artificial systems have become particularly important. Due to its ethical (Metzinger, 2021), legal and societal relevance, the question of whether artificial systems are or can be conscious, referred to as the question of *synthetic phenomenology* or *artificial consciousness*, is in need of particular attention.

Because AI systems are mathematical systems—they are defined by formal or mathematical structures, both on the level of programming and the level of machine code—the question of synthetic phenomenology is particularly amenable to mathematical tools. To apply a theory or concept to an AI system, the theory or concept needs to be flashed out in formal details. Hence, artificial consciousness has become a major topic of interest in mathematical consciousness science (Association for Mathematical Consciousness Science, 2023). As part of this PhD, two lines of research have been pursued that target synthetic phenomenology.

1.4.1. Implications of CPU and GPU Design

The first line of research, carried out by Kleiner and Ludwig (in press), is presented in **Chapter 10**. It rests on an analysis of the central component of AI systems: their processing units.

Contemporary AI systems, for example Generative Pre-trained Transformers (GPTs), which include Large Language Models (LLMs), are computer programmes. They consist of a few hundred lines of code (almost nothing compared to the tens of millions of lines of code of operating systems like Windows, macOS, or Linux) and a large file of several hundred gigabytes which only contains numbers.[6] What brings these two things together are processing units (PUs). The numbers are converted to strings of zeros and ones, and the code file, once compiled and executed, instructs the processing unit what to do with these strings. If one runs an AI on one's own computer, the task is done by one's central processing unit (CPU), but more advanced systems usually make use of graphics processing units (GPUs), tensor processing units (TPUs), or, as of late, data processing units (DPUs). All of them crunch the numbers as specified in the code, but they differ in how optimised and effective they are in doing this. Because of this, as far as the physical substrate is concerned, contemporary AI systems actually are processing units (PUs). PUs are what supports an AI, much like brains are what supports you.

In light of this it is very surprising that the nature of PUs has not received any attention in investigations of whether AIs have minds, including questions of AI consciousness, prior to the work carried out in (Kleiner & Ludwig, in press). That is the case even though PUs are fundamentally different from brains and biological substrate.

[6] The numbers are the weights of an artificial neural network, which result from a training task that makes use of a large part of the internet. Training is the difficult and expensive part of creating an LLM. Running the LLM is comparably cheap and can, with enough patience, also be done on a personal computer.

The largest difference between biological systems like brains and PUs is that PUs are designed and verified to behave exactly as specified by a formal system in the sense of mathematics: the calculation in terms of zeros and ones on the chip is precisely governed by pre-set mathematical rules. "Artificial", in this case not only means man-made, but it means that the system is made to behave in an exact pre-specified way.

The analysis in Chapter 10 shows that this fact has strong consequences if consciousness is dynamically relevant. Intuitively speaking, this is the case because dynamical relevance requires consciousness to be able to make some difference in its substrate, e.g. in the case of a system's report about its conscious experience. But the exact adherence of a PU to a pre-set formal system ensures this can't happen.

The result in Chapter 10 is presented in the form of a no-go theorem, which establishes the conclusion based on a formal proof. No-go theorems serve an important role in scientific progress in physics, and we discuss how they can contribute to consciousness science in Section 12.5 of the conclusion.

1.4.2. Mortal Computation

The second line of research regarding artificial consciousness pursued in this PhD is presented in **Chapter 11**. It also focuses on the distinction between biological and artificial systems, though not on the level of substrate, as in Chapter 10, but rather on the level of computation.

The idea—or better: the observation—that there is a difference between the computation that computers implement, and the computations that biological systems like brains implement, called *neural computation* (Piccinini, 2020), is not new. There are various differences between PUs and brains, and these differences are reflected in models of computation that these systems can instantiate.

There is, however, a deeper difference that goes beyond questions of implementation. This difference was first observed by Hinton (2022), and is called *mortal computation*.

In a nutshell, a computation is mortal if it cannot be separated from the hardware on which it runs; if "it dies with the hardware" (Hinton, 2022). All computations carried out by computers to date are immortal, they can be separated from the hardware. In contrast, computations carried out by biological systems are mortal, they cannot be separated from the hardware, because biological computation, which is learned rather than programmed, relies on "large and unknown variations in the connectivity and non-linearities of different instances of hardware" (ibid.). Even if it were possible to copy a mortal computation to another system, it would cease to work.

Chapter 11 is a first indication that consciousness may require mortal computation. The chapter shows that computational functionalism—the very idea that consciousness is a computation—implies that consciousness is a mortal computation. That is surprising because the 'computations' in computational functionalism are often conceived of as being Turing computations, examples of which are the programs we run on today's computers and mobile devices. Therefore, the result runs counter to many intuitions. But it is aligned with the original definition of computational functionalism by Putnam

(1967), which makes use of probabilistic automata descriptions rather than Turing machines, and considers biological organisms as examples. The result is also surprising because it shows that if computational functionalism were indeed true, then contemporary and near-future AI systems, which are immortal computations, could not be conscious, contrary to thinking in several contemporary analyses, like (Butlin et al., 2023).

It should be emphasised, though, that neither of the results presented in Chapter 10 and 11 attempt to provide a final answer to the question of artificial consciousness. That is the case because they both only target contemporary and near future systems: contemporary and near-future processing units in the case of Chapter 10, and contemporary and near-future forms of computation in the case of Chapter 11. New developments in the semiconductor industry, for example regarding analogue computations, indicate a trend towards transcending both.

1.5. Mathematics in Consciousness Science

No natural science has, so far, been solved without mathematics. And consciousness is a natural phenomenon. Hence it is no surprise that as consciousness science starts to blossom, mathematical questions, problems and tasks come to surface as well.

A large part of the agenda of those who work on mathematical questions in consciousness science is to support the development of experiments and theories. Mathematical consciousness science is, to a large extent, a service department to other branches of consciousness science. Several projects in this PhD can be understood in this way, from the mathematical analyses of IIT and Predictive Processing in Chapters 2, 3 and 4, the analysis and modelling of experiments in Chapters 5 and 6, the discussion of contemporary thinking on structuralist research in Chapter 7, and the contributions to the debate on AI consciousness in Chapters 10 and 11.

But as is exemplified by mathematical physics, mathematical approaches may also, occasionally, make contributions that attempt to push the boundaries of scientific progress themselves. Within this PhD, if anywhere, this has been the case in the work on foundations of structural approaches in Chapters 8 and 9. This work is not finished, of course, but the hope is that a few first steps into the right direction have been taken.

This chapter has surveyed the past. In Chapter 12, we attempt to provide an outlook into the future. We will sketch how the research carried out in this PhD could be pursued further, and which opportunities this might afford for consciousness science at large.

Part I.

On Theories

2. The Mathematical Structure of Integrated Information Theory

Johannes Kleiner, Sean Tull[1]

2.1. Introduction

Integrated Information Theory (IIT), developed by Giulio Tononi and collaborators (Tononi, 2004, 2005, 2008; Balduzzi & Tononi, 2008), has emerged as one of the leading scientific theories of consciousness. At the heart of the latest version of the theory (Oizumi et al., 2014; Marshall, Gomez-Ramirez, & Tononi, 2016; Tononi, Boly, Massimini, & Koch, 2016; Mayner et al., 2018; C. Koch, Massimini, Boly, & Tononi, 2016) is an algorithm which, based on the level of *integration* of the internal functional relationships of a physical system in a given state, aims to determine both the quality and quantity ('Φ value') of its conscious experience.

While promising in itself (A. M. Haun et al., 2016; Tsuchiya, Haun, Cohen, & Oizumi, 2016), the mathematical formulation of the theory is not satisfying to date. The presentation in terms of examples and accompanying explanation veils the essential mathematical structure of the theory and impedes philosophical and scientific analysis. In addition, the current definition of the theory can only be applied to comparably simple classical physical systems (A. B. Barrett, 2014), which is problematic if the theory is taken to be a fundamental theory of consciousness, and should eventually be reconciled with our present theories of physics.

[1]Published as: Kleiner, J., & Tull, S. (2021). The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, 6, 602973. (Kleiner & Tull, 2021)

To resolve these problems, we examine the essentials of the IIT algorithm and formally define a generalized notion of Integrated Information Theory. This notion captures the inherent mathematical structure of IIT and offers a rigorous mathematical definition of the theory which has 'classical' IIT 3.0 of Tononi et al. (Oizumi et al., 2014; Marshall et al., 2016; Mayner et al., 2018) as well as the more recently introduced *Quantum Integrated Information Theory* of Zanardi, Tomka and Venuti (Zanardi, Tomka, & Venuti, 2018) as special cases. In addition, this generalization allows us to extend classical IIT, freeing it from a number of simplifying assumptions identified in (A. B. Barrett & Mediano, 2019).

This work is concerned with the most recent version of IIT as proposed in (Oizumi et al., 2014; Marshall et al., 2016; Tononi et al., 2016; Mayner et al., 2018) and similar papers quoted below. Thus our constructions recover the specific theory of consciousness referred to as IIT 3.0 or IIT 3.x. Earlier proposals by Tononi et al. that also aim to explicate the general idea of an essential connection between consciousness and integrated information constitute alternative theories of consciousness which we do not study here. A yet different approach would be to take the term 'Integrated Information Theory' to refer to the general idea of associating conscious experience with some pre-theoretic notion of integrated information, and to explore the different ways of how this notion could be defined in formal terms (A. B. Barrett & Seth, 2011; Seth, Barrett, & Barnett, 2011; P. A. Mediano, Rosas, Carhart-Harris, Seth, & Barrett, 2019; P. A. Mediano, Seth, & Barrett, 2019).

In the associated article (Tull & Kleiner, 2021) we show more generally how the main notions of IIT, including causation and integration, can be treated, and an IIT defined, starting from any suitable theory of physical systems and processes described in terms of category theory. Restricting to classical or quantum process then yields each of the above as special cases. This treatment makes IIT applicable to a large class of physical systems and helps overcome the current restrictions.

Our definition of IIT may serve as the starting point for further mathematical analysis of IIT, in particular if related to category theory (Tsuchiya, Taguchi, & Saigo, 2016; Northoff, Tsuchiya, & Saigo, 2019). It also provides a simplification and mathematical clarification of the IIT algorithm which extends the technical analysis of the theory (A. B. Barrett, 2014; Tegmark, 2015, 2016) and may contribute to its ongoing critical discussion (A. B. Barrett & Seth, 2011; Peressini, 2013; Cerullo, 2015; Bayne, 2018; P. A. Mediano, Seth, & Barrett, 2019; P. A. Mediano, Rosas, et al., 2019; McQueen, 2019). The concise presentation of IIT in this article should also help to make IIT more easily accessible for mathematicians, physicists and other researchers with a strongly formal background.

Relation to other work This work develops a thorough mathematical perspective of one of the promising contemporary theories of consciousness. As such it is part of a number of recent contributions which seek to explore the role and prospects of mathematical theories of consciousness (Tsuchiya, Taguchi, & Saigo, 2016; Hardy, 2017; Kent, 2018; Northoff et al., 2019; Kleiner, 2020b), to help overcome problems of existing mod-

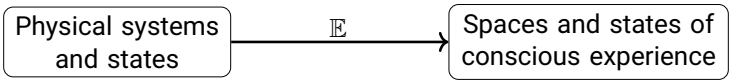


Figure 2.1.: An Integrated Information Theory specifies for every system in a particular state its conscious experience, described formally as an element of an experience space. In our formalization, this is a map

$$\mathbf{Sys} \xrightarrow{\mathbb{E}} \mathbf{Exp}$$

from the system class $\mathbf{Sys}$ into a class $\mathbf{Exp}$ of experience spaces, which, first, sends each system S to its space of possible experiences $\mathbb{E}(S)$, and, second, sends each state $s \in \mathsf{St}(S)$ to the actual experience the system is having when in that space,

$$\mathsf{St}(S) \to \mathbb{E}(S) \qquad s \mapsto \mathbb{E}(S, s)\ .$$

The definition of this map in terms of axiomatic descriptions of physical systems, experience spaces and further structure used in classical IIT is given in the first half of this paper.

els (Resende, 2018; Kleiner, 2020b; Kleiner & Hoel, 2021) and to eventually develop new proposals (Chang, Biehl, Yu, & Kanai, 2020; Kent, 2020; Mason, 2016; Kremnizer & Ranchin, 2015; Hoffman & Prakash, 2014; Mueller, 2017; Signorelli, Wang, & Khan, 2021).

2.1.1. Structure of article

We begin by introducing the necessary ingredients of a generalised Integrated Information Theory in Sections 2.2 to 2.4, namely physical systems, spaces of conscious experience and cause-effect repertoires. Our approach is *axiomatic* in that we state only the precise formal structure which is necessary to apply the IIT algorithm. We neither motivate nor criticize these structures as necessary or suitable to model consciousness. Our goal is simply to recover IIT 3.0. In Section 2.5, we introduce a simple formal tool which allows us to present the definition of the algorithm of an IIT in a concise form in Sections 2.6 and 2.7. Finally, in Section 2.8, we summarise the full definition of such a theory.

Following this we give several examples including IIT 3.0 in Section 2.9 and Quantum IIT in Section 2.10. In Section 2.11 we discuss how our formulation allows one to extend classical IIT in several fundamental ways, before discussing further modifications to our approach and other future work in Section 2.12. Finally, the appendix includes a detailed explanation of how our generalization of IIT coincides with its usual presentation in the case of classical IIT.

2.2. Systems

The first step in defining an Integrated Information Theory (IIT) is to specify a class $\mathbf{Sys}$ of physical *systems* to be studied. Each element $S \in \mathbf{Sys}$ is interpreted as a model of one particular physical system. In order to apply the IIT algorithm, it is only necessary that each element S come with the following features.

Definition 2.2.1. *A* system class $\mathbf{Sys}$ *is a class each of whose elements* S*, called* systems*, come with the following data:*

1. *a set* $\mathit{St}(S)$ *of* states*;*
2. *for every* $s \in \mathit{St}(S)$*, a set* $\mathrm{Sub}_s(S) \subset \mathbf{Sys}$ *of subsystems and for each* $M \in \mathrm{Sub}_s(S)$ *an induced state* $s|_M \in \mathit{St}(M)$*;*
3. *a set* $\mathbb{D}_S$ *of* decompositions*, with a given* trivial decomposition $1 \in \mathbb{D}_S$*;*
4. *for each* $z \in \mathbb{D}_S$ *a corresponding* cut system $S^z \in \mathbf{Sys}$ *and for each state* $s \in \mathit{St}(S)$ *a corresponding* cut state $s^z \in \mathit{St}(S^z)$.

Moreover, we require that $\mathbf{Sys}$ contains a distinguished *empty system*, denoted I, and that $I \in \mathrm{Sub}(S)$ for all S. For the IIT algorithm to work, we need to assume furthermore that the number of subsystems remains the same under cuts or changes of states, i.e. that we have bijections $\mathrm{Sub}_s(S) \simeq \mathrm{Sub}_{s'}(S)$ for all $s, s' \in \mathrm{St}(S)$ and $\mathrm{Sub}_s(S) \simeq \mathrm{Sub}_{s^z}(S^z)$ for all $z \in \mathbb{D}_S$.

Note that taking a subsystem of a system S requires specifying a state s of S. An example class of systems is illustrated in Figure 2.2. In this article we will assume that each set $\mathrm{Sub}_s(S)$ is finite, discussing the extension to the infinite case in Section 2.12. We will give examples of system classes and for all following definitions in Sections 2.9 and 2.10.

2.3. Experience

An IIT aims to specify for each system in a particular state its *conscious experience*. As such, it will require a mathematical model of such experiences. Examining classical IIT, we find the following basic features of the final experiential states it describes which are needed for its algorithm.

Firstly, each experience e should crucially come with an *intensity*, given by a number $\| e \|$ in the non-negative reals $\mathbb{R}^+$ (including zero). This intensity will finally correspond to the overall intensity of experience, usually denoted by Φ. Next, in order to compare experiences, we require a notion of *distance* $d(e, e')$ between any pair of experiences e, e'. Finally, the algorithm will require us to be able to *rescale* any given experience e to have any given intensity. Mathematically, this is most easily encoded by letting us multiply any experience e by any number $r \in \mathbb{R}^+$. In summary, a minimal model of experience in a generalized IIT is the following.

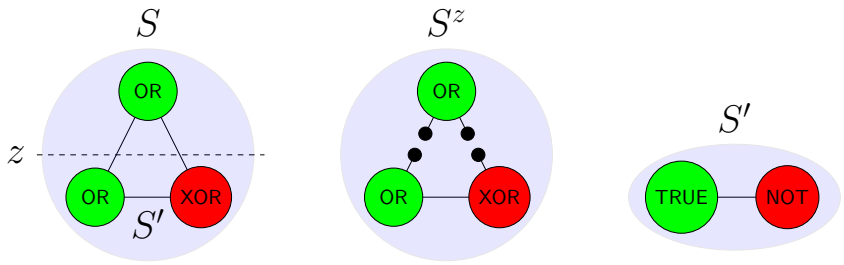


Figure 2.2.: As an example of Definition 2.2.1 similar to IIT 3.0, consider simple systems given by sets of nodes (or 'elements'), with a state assigning each node the state 'on' (depicted green) or 'off' (red). Each system comes with a time evolution shown by writing on each node how its state in the next time-step depends on the states of the others now.

Decompositions of a system S correspond to binary partition of the nodes, such as z above. The cut system S^z is given by modifying the time evolution of S so that the two halves do not interact; in this case all connections between the halves are replaced by sources of noise which send 'on' or 'off' with equal likelihood, depicted as black dots above.
Given a current state s of S, any subset of the nodes (such as those below the dotted line) determines a subsystem S', with time evolution obtained from that of S by fixing the nodes in $S \setminus S'$ (here, the upper node) to be in the state specified by s.
Note that while in this example any subsystem (subset of S) determines a decomposition (partition of S) we do not require such a relationship in general.

Definition 2.3.1. *An* experience space *is a set E with:*

1. *an* intensity *function* $\|\,.\,\|\colon E \to \mathbb{R}^+$*;*
2. *a* distance *function* $d\colon E \times E \to \mathbb{R}^+$*;*
3. *a* scalar multiplication $\mathbb{R}^+ \times E \to E$*, denoted* $(r, e) \mapsto r \cdot e$*, satisfying*

$$\|\, r \cdot e \,\| = r \cdot \|\, e \,\| \qquad r \cdot (s \cdot e) = (rs) \cdot e \qquad 1 \cdot e = e$$

for all $e \in E$ and $r, s \in \mathbb{R}^+$.

We remark that this same axiomatisation will apply both to the full space of experiences of a system, as well as to the the spaces describing components of the experiences ('concepts' and 'proto-experiences' defined in later sections). We note that the distance function does not necessarily have to satisfy the axioms of a metric. While this and further natural axioms such as $d(r \cdot e, r \cdot f) = r \cdot d(e, f)$ might hold, they are not necessary for the IIT algorithm.

The above definition is very general, and in any specific application of IIT, the experiences may come with further mathematical structure. The following example describes the experience space used in classical IIT.

Example 2.3.2. *Any metric space (X, d) may be extended to an experience space $\bar{X} := X \times \mathbb{R}^+$ in various ways. E.g., one can define $\|(x, r)\| = r$, $r \cdot (x, s) = (x, rs)$ and define the distance as*

$$d\big((x, r), (y, s)\big) = r\, d(x, y) \tag{2.1}$$

This is the definition used in classical IIT (cf. Section 2.9 and Appendix 2.A).

An important operation on experience spaces is taking their *product*.

Definition 2.3.3. *For experience spaces E and F, we define the product to be the space $E \times F$ with distance*

$$d\big((e, f), (e', f')\big) = d(e, e') + d(f, f')\,, \tag{2.2}$$

intensity $\|(e, f)\| = \max\{\|e\|, \|f\|\}$ and scalar multiplication $r \cdot (e, f) = (r \cdot e, r \cdot f)$. This generalizes to any finite product $\prod_{i \in I} E_i$ of experience spaces.

2.4. Repertoires

In order to define the experience space and individual experiences of a system S, an IIT utilizes basic building blocks called 'repertoires', which we will now define. Next to the specification of a system class, this is the essential data necessary for the IIT algorithm to be applied.

Each repertoire describes a way of 'decomposing' experiences, in the following sense. Let D denote any set with a distinguished element 1, for example the set $\mathbb{D}_S$ of decompositions of a system S, where the distinguished element is the trivial decomposition $1 \in \mathbb{D}_S$.

Definition 2.4.1. *Let e be an element of an experience space E. A* decomposition of e over D *is a mapping $\bar{e}\colon D \to E$ with $\bar{e}(1) = e$.*

In more detail, a repertoire specifies a proto-experience for every pair of subsystems and describes how this experience changes if the subsystems are decomposed. This allows one to assess how integrated the system is with respect to a particular repertoire. Two repertoires are necessary for the IIT algorithm to be applied, together called the cause-effect repertoire.

For subsystems $M, P \in \mathrm{Sub}_s(S)$, define $\mathbb{D}_{M,P} := \mathbb{D}_M \times \mathbb{D}_P$. This set describes the decomposition of both subsystems simultaneously. It has a distinguished element $1 = (1_M, 1_P)$.

Definition 2.4.2. *A* cause-effect repertoire *at S is given by a choice of experience space $\mathbb{PE}(S)$, called the space of* proto-experiences*, and for each $s \in \mathit{St}(S)$ and $M, P \in \mathrm{Sub}_s(S)$, a pair of elements*

$$\mathsf{caus}_s(M, P)\,,\ \mathsf{eff}_s(M, P)\ \in\ \mathbb{PE}(S) \tag{2.3}$$

and for each of them a decomposition over $\mathbb{D}_{M,P}$.

Examples of cause-effect repertoires will be given in Sections 2.9 and 2.10. A general definition in terms of process theories is given in (Tull & Kleiner, 2021). For the IIT algorithm, a cause-effect repertoire needs to be specified for every system S, as in the following definition.

Definition 2.4.3. *A* cause-effect structure *is a specification of a cause-effect repertoire for every $S \in \mathbf{Sys}$ such that*

$$\mathbb{PE}(S) = \mathbb{PE}(S^z) \quad \textit{for all} \quad z \in \mathbb{D}_S \,. \tag{2.4}$$

The names 'cause' and 'effect' highlight that the definitions of $\mathsf{caus}_s(M, P)$ and $\mathsf{eff}_s(M, P)$ in classical and quantum IIT describe the causal dynamics of the system. More precisely, they are intended to capture the manner in which the 'current' state s of the system, when restricted to M, constrains the 'previous' or 'next' state of P, respectively.

2.5. Integration

We have now introduced all of the data required to define an IIT; namely, a system class along with a cause-effect structure. From this, we will give an algorithm aiming to specify the conscious experience of a system. Before proceeding to do so, we introduce a conceptual short-cut which allows the algorithm to be stated in a concise form. This captures the core ingredient of an IIT, namely the computation of how integrated an entity is.

Definition 2.5.1. *Let E be an experience space and e an element with a decomposition over some set D. The* integration level *of e relative to this decomposition is*

$$\phi(e) := \min_{1 \neq z \in D} d(e, \bar{e}(z)) \,. \tag{2.5}$$

Here, d denotes the distance function of E, and the minimum is taken over all elements of D besides 1. The integration scaling *of e is then the element of E defined by*

$$\iota(e) := \phi(e) \cdot \hat{e} \,, \tag{2.6}$$

where $\hat{e}$ denotes the normalization *of e, defined as*

$$\hat{e} := \begin{cases} \frac{1}{\| e \|} \cdot e & \textit{if } \| e \| \neq 0 \\ e & \textit{if } \| e \| = 0 \,. \end{cases}$$

Finally, the integration scaling *of a pair e_1, e_2 of such elements is the pair*

$$\iota(e_1, e_2) := (\phi \cdot \hat{e_1}, \phi \cdot \hat{e_2}) \tag{2.7}$$

where $\phi := \min(\phi(e_1), \phi(e_2))$ is the minimum of their integration levels.

We will also need to consider indexed collections of decomposable elements. Let S be a system in a state $s \in \mathrm{St}(S)$ and assume that for every $M \in \mathrm{Sub}_s(S)$ an element e_M of some experience space E_M with a decomposition over some D_M is given. We call $(e_M)_{M \in \mathrm{Sub}_s(S)}$ a *collection* of decomposable elements, and denote it as $(e_M)_M$.

Definition 2.5.2. *The* core *of the collection $(e_M)_M$ is the subsystem $C \in \mathrm{Sub}(S)$ for which $\phi(e_C)$ is maximal.*[2] *The* core integration scaling *of the collection is $\iota(e_C)$. The* core integration scaling *of a pair of collections $(e_M, f_M)_M$ is $\iota(e_C, f_D)$, where C, D are the cores of $(e_M)_M$ and $(f_M)_M$, respectively.*

2.6. Constructions - Mechanism Level

Let $S \in \mathbf{Sys}$ be a physical system whose experience in a state $s \in \mathrm{St}(S)$ is to be determined. The first level of the algorithm involves fixing some subsystem $M \in \mathrm{Sub}_s(S)$, referred to as a 'mechanism', and associating to it an object called its 'concept' which belongs to the *concept space*

$$\mathbb{C}(S) := \mathbb{PE}(S) \times \mathbb{PE}(S) \,. \tag{2.8}$$

For every choice of $P \in \mathrm{Sub}_s(S)$, called a 'purview', the repertoire values $\mathsf{caus}_s(M, P)$ and $\mathsf{eff}_s(M, P)$ are elements of $\mathbb{PE}(S)$ with given decompositions over $\mathbb{D}_{M,P}$. Fixing M, they form collection of decomposable elements,

$$\begin{aligned} \mathsf{caus}_s(M) &:= (\mathsf{caus}_s(M, P))_{P \in \mathrm{Sub}(S)} \\ \mathsf{eff}_s(M) &:= (\mathsf{eff}_s(M, P))_{P \in \mathrm{Sub}(S)} \,. \end{aligned} \tag{2.9}$$

The *concept* of M is then defined as the core integration scaling of this pair of collections,

$$\mathbb{C}_{S,s}(M) := \text{Core integration scaling of } (\mathsf{caus}_s(M), \mathsf{eff}_s(M)) \,. \tag{2.10}$$

It is an element of $\mathbb{C}(S)$. Unravelling our definitions, the concept thus consists of the values of the cause and effect repertoires at their respective 'core' purviews P^c, P^e, i.e. those which make them 'most integrated'. These values $\mathsf{caus}(M, P^c)$ and $\mathsf{eff}(M, P^e)$ are then each rescaled to have intensity given by the minima of their two integration levels.

2.7. Constructions - System Level

The second level of the algorithm specifies the experience of the system S in state s. To this end, all concepts of a system are collected to form its *Q-shape*, defined as

$$\mathbb{Q}_s(S) := (\mathbb{C}_{S,s}(M))_{M \in \mathrm{Sub}_s(S)} \,. \tag{2.11}$$

[2] If the maximum does not exist, we define the core to be the empty system I.

This is an element of the space

$$\mathbb{E}(S) = \mathbb{C}(S)^{n(S)}\,, \tag{2.12}$$

where $n(S) := |\mathrm{Sub}_s(S)|$, which is finite and independent of the state s according to our assumptions. We can also define a Q-shape for any cut of S. Let $z \in \mathbb{D}_S$ be a decomposition, S^z the corresponding cut system and s^z be the corresponding cut state. We define

$$\mathbb{Q}_s(S^z) := (\mathbb{C}_{S^z,s^z}(M))_{M\in\mathrm{Sub}_{s^z}(S^z)}\,. \tag{2.13}$$

Because of (2.4), and since the number of subsystems remains the same when cutting, $\mathbb{Q}_s(S^z)$ is also an element of $\mathbb{E}(S)$. This gives a map

$$\begin{aligned}\bar{\mathbb{Q}}_{S,s} : \mathbb{D}_S &\to \mathbb{E}(S)\\ z &\mapsto \mathbb{Q}_s(S^z)\end{aligned}$$

which is a decomposition of $\mathbb{Q}_s(S)$ over $\mathbb{D}_S$. Considering this map for every subsystem of S gives a collection of decompositions defined as

$$\mathbb{Q}(S,s) := \left(\bar{\mathbb{Q}}_{M,s|_M}\right)_{M\in\mathrm{Sub}_s(S)}$$

This is the system level-object of relevance and is what specifies the experience of a system according to IIT.

Definition 2.7.1. *The actual* experience *of the system S in the state $s \in St(S)$ is*

$$\mathbb{E}(S,s) := \mathit{Core\ integration\ scaling\ of}\ \mathbb{Q}(S,s)\,. \tag{2.14}$$

The definition implies that $\mathbb{E}(S,s) \in \mathbb{E}(M)$, where $M \in \mathrm{Sub}_s(S)$ is the core of the collection $\mathbb{Q}(S,s)$, called the *major complex*. It describes which part of the system S is actually conscious. In most cases there will be a natural embedding $\mathbb{E}(M) \hookrightarrow \mathbb{E}(S)$ for a subsystem M of S, allowing us to view $\mathbb{E}(S,s)$ as an element of $\mathbb{E}(S)$ itself. Assuming this embedding to exist allows us to define an Integrated Information Theory concisely in the next section.

2.8. Integrated Information Theories

We can now summarize all that we have said about IITs.

Definition 2.8.1. *An* Integrated Information Theory *is determined as follows. The* data *of the theory is a system class $\mathbf{Sys}$ along with a cause-effect structure. The theory then gives a mapping*

$$\mathbf{Sys} \xrightarrow{\mathbb{E}} \mathbf{Exp} \tag{2.15}$$

into the class $\mathbf{Exp}$ *of all experience spaces, sending each system* S *to its space of experiences* $\mathbb{E}(S)$ *defined in* (2.12)*, and a mapping*

$$\begin{aligned} \mathit{St}(S) &\to \mathbb{E}(S) \\ s &\mapsto \mathbb{E}(S, s) \end{aligned} \tag{2.16}$$

which determines the experience of the system when in a state s*, defined in* (2.14)*. The* quantity *of the system's experience is given by*

$$\Phi(S, s) := \| \mathbb{E}(S, s) \| ,$$

and the quality of the system's experience is given by the normalized experience $\hat{\mathbb{E}}(S, s)$*. The experience is located in the core of the collection* $\mathbb{Q}(S, s)$*, called* major complex*, which is a subsystem of* S*.*

In the next sections we specify the data of several example IITs.

2.9. Classical IIT

In this section we show how IIT 3.0 (Mayner et al., 2018; Marshall et al., 2016; Tononi, 2015; Oizumi et al., 2014) fits in into the framework developed here. A detailed explanation of how our earlier algorithm fits with the usual presentation of IIT is given in Appendix 2.A. In (Tull & Kleiner, 2021) we give an alternative categorical presentation of the theory.

2.9.1. Systems

We first describe the system class underlying classical IIT. Physical systems S are considered to be built up of several components $S_1, \dots, S_n$, called *elements*. Each element S_i comes with a finite set of states $\mathsf{St}(S_i)$, equipped with a metric. A state of S is given by specifying a state of each element, so that

$$\mathsf{St}(S) = \mathsf{St}(S_1) \times ... \times \mathsf{St}(S_n) . \tag{2.17}$$

We define a metric d on $\mathsf{St}(S)$ by summing over the metrics of the element state spaces $\mathsf{St}(S_i)$ and denote the collection of probability distributions over $\mathsf{St}(S)$ by $\mathcal{P}(S)$. Note that we may view $\mathsf{St}(S)$ as a subset of $\mathcal{P}(S)$ by identifying any $s \in \mathsf{St}(S)$ with its Dirac distribution $\delta_s \in \mathcal{P}(S)$, which is why we abbreviate δ_s by s occasionally in what follows.

Additionally, each system comes with a probabilistic (discrete) *time evolution operator* or *transition probability matrix*, sending each $s \in \mathsf{St}(S)$ to a probabilistic state $T(s) \in \mathcal{P}(S)$. Equivalently it may be described as a convex-linear map

$$T \colon \mathcal{P}(S) \to \mathcal{P}(S) . \tag{2.18}$$

Furthermore, the evolution T is required to satisfy a property called *conditional independence*, which we define shortly.

The class $\mathbf{Sys}$ consists of all possible tuples $S = (\{S_i\}_{i=1}^n, T)$ of this kind, with the trivial system I having only a single element with a single state and trivial time evolution.

2.9.2. Conditioning and Marginalizing

In what follows, we will need to consider two operations on the map T. Let M be any subset of the elements of a system and $M^\perp$ its complement. We again denote by $\mathrm{St}(M)$ the Cartesian product of the states of all elements in M, and by $\mathcal{P}(M)$ the probability distributions on $\mathrm{St}(M)$. For any $p \in \mathcal{P}(M)$, we define the *conditioning* (Mayner et al., 2018) of T on p as the map

$$\begin{aligned} T_{|p\rangle} \colon \mathcal{P}(M^\perp) &\to \mathcal{P}(S) \\ p' &\mapsto T(p \cdot p') \end{aligned} \tag{2.19}$$

where $p \cdot p'$ denotes the multiplication of these probability distributions to give a probability distribution over S. Next, we define *marginalisation over* M as the map

$$\langle M| \colon \mathcal{P}(S) \to \mathcal{P}(M^\perp) \tag{2.20}$$

such that for each $p \in \mathcal{P}(S)$ and $m_2 \in \mathrm{St}(M^\perp)$ we have

$$\langle M|(p)(m_2) = \sum_{m_1 \in \mathrm{St}(M)} p(m_1, m_2)\,. \tag{2.21}$$

In particular for any map T as above we call $\langle M|T$ the *marginal* of T over M and we write $T_i := \langle S_i^\perp|T$ for each $i = 1, \ldots, n$. Conditional independence of T may now be defined as the requirement that

$$T(p) = \prod_{i=1}^{n} T_i(p) \qquad \text{for all } p \in \mathcal{P}(S)\,,$$

where the right-hand side is again a probability distribution over $\mathrm{St}(S)$.

2.9.3. Subsystems, Decompositions and Cuts

Let a system S in a state $s \in \mathrm{St}(S)$ be given. The subsystems are characterized by subsets of the elements that constitute S. For any subset $M = \{S_1, ..., S_m\}$ of the elements of S, the corresponding subsystem is also denoted M and $\mathrm{St}(M)$ is again given by the product of the $\mathrm{St}(S_i)$, with time evolution

$$T_M := \langle M^\perp|T|s_{M^\perp}\rangle\,, \tag{2.22}$$

where $s_{M^\perp}$ is the restriction of the state s to $\mathrm{St}(M^\perp)$ and $|s_{M^\perp}\rangle$ denotes the conditioning on the Dirac distribution $\delta_{s_{M^\perp}}$.

The decomposition set $\mathbb{D}_S$ of a system S consists of all partitions of the set N of elements of S into two disjoint sets M and $M^\perp$. We denote such a partition by $z = (M, M^\perp)$. The trivial decomposition 1 is the pair $(N, \emptyset)$.

For any decomposition $(M, M^\perp)$ the corresponding cut system $S^{(M,M^\perp)}$ is the same as S but with a new time evolution $T^{(M,M^\perp)}$. Using conditional independence, it may be defined for each $i = 1, \ldots, n$ as

$$T_i^{(M,M^\perp)} := \begin{cases} T_i & i \in M^\perp \\ T_i|\omega_{M^\perp}\rangle\langle M^\perp| & i \in M \end{cases} \tag{2.23}$$

where $\omega_M \in \mathcal{P}(M)$ denotes the uniform distribution on $\mathrm{St}(M)$. This is interpreted in the graph depiction as removing all those edges from the graph whose source is in $M^\perp$ and whose target is in M. The corresponding input of the target element is replaced by noise, i.e. the uniform probability distribution over the source element.

2.9.4. Proto-Experiences

For each system S, the first Wasserstein metric (or 'Earth Mover's Distance') makes $\mathcal{P}(S)$ a metric space $(\mathcal{P}(S), d)$. The space of proto-experiences of classical IIT is

$$\mathbb{PE}(S) := \overline{\mathcal{P}(S)}\,, \tag{2.24}$$

where $\overline{\mathcal{P}(S)}$ is defined in Example 2.3.2. Thus elements of $\mathbb{PE}(S)$ are of the form (p, r) for some $p \in \mathcal{P}(S)$ and $r \in \mathbb{R}^+$, with distance function, intensity and scalar multiplication as defined in the example.

2.9.5. Repertoires

It remains to define the cause-effect repertoires. Fixing a state s of S, the first step will be to define maps caus'_s and eff'_s which send any choice of $(M, P) \in \mathrm{Sub}(S) \times \mathrm{Sub}(S)$ to an element of $\mathcal{P}(P)$. These should describe the way in which the current state of M constrains that of P in the next or previous time-steps. We begin with the effect repertoire. For a single element purview P_i we define

$$\mathsf{eff}'_s(M, P_i) := \langle P_i^\perp|T|\omega_{M^\perp}\rangle(s_M)\,, \tag{2.25}$$

where s_M denotes (the Dirac distribution of) the restriction of the state s to M. While it is natural to use the same definition for arbitrary purviews, IIT 3.0 in fact uses another definition based on consideration of 'virtual elements' (Mayner et al., 2018; Marshall et al., 2016; Tononi, 2015), which also makes calculations more efficient (Mayner et al., 2018, Supplement S1). For general purviews P, this definition is

$$\mathsf{eff}'_s(M, P) = \prod_{P_i \in P} \mathsf{eff}'_s(M, P_i)\,, \tag{2.26}$$

taking the product over all elements P_i in the purview P. Next, for the cause repertoire, for a single element mechanism M_i and each $\tilde{s} \in \mathrm{St}(P)$, we define

$$\mathsf{caus}'_s(M_i, P)[\tilde{s}] = \lambda\, \langle M_i^\perp|T|\omega_{P^\perp}\rangle(\delta_{\tilde{s}})[s_{M_i}]\,, \tag{2.27}$$

where λ is the unique normalisation scalar making $\mathsf{caus}'_s(M_i, P)$ a valid element of $\mathcal{P}(P)$. Here, for clarity, we have indicated evaluation of probability distributions at particular states by square brackets. If the time evolution operator has an inverse T^{-1}, this cause repertoire could be defined similarly to (2.25) by $\mathsf{caus}'_s(M_i, P) = \langle P^\perp | T^{-1} | \omega_{M_i^\perp} \rangle (s_{M_i})$, but classical IIT does not utilize this definition.

For general mechanisms M, we then define

$$\mathsf{caus}'_s(M, P) = \kappa \prod_{M_i \in M} \mathsf{caus}'_s(M_i, P) \tag{2.28}$$

where the product is over all elements M_i in M and where $\kappa \in \mathbb{R}^+$ is again a normalisation constant. We may at last now define

$$\begin{aligned} \mathsf{caus}_s(M, P) &:= \mathsf{caus}'_s(M, P) \cdot \mathsf{caus}'_s(\emptyset, P^\perp) \\ \mathsf{eff}_s(M, P) &:= \mathsf{eff}'_s(M, P) \cdot \mathsf{eff}'_s(\emptyset, P^\perp) \,, \end{aligned} \tag{2.29}$$

with intensity 1 when viewed as elements of $\mathbb{PE}(S)$. Here, the dot indicates again the multiplication of probability distributions and $\emptyset$ denotes the empty mechanism.

The distributions $\mathsf{caus}'_s(\emptyset, P^\perp)$ and $\mathsf{eff}'_s(\emptyset, P^\perp)$ are called the *unconstrained cause and effect repertoires* over $P^\perp$.

Remark 2.9.1. *It is in fact possible for the right-hand side of* (2.27) *to be equal to* 0 *for all* $\tilde{s}$ *for some* $M_i \in M$*. In this case we set* $\mathsf{caus}_s(M, P) = (\omega_S, 0)$ *in* $\mathbb{PE}(S)$*.*

Finally we must specify the decompositions of these elements over $\mathbb{D}_{M,P}$. For any partitions $z_M = (M_1, M_2)$ of M and $z_P = (P_1, P_2)$ of P, we define

$$\begin{aligned} \overline{\mathsf{caus}_s}(M, P)(z_M, z_P) &:= \mathsf{caus}'_s(M_1, P_1) \cdot \mathsf{caus}'_s(M_2, P_2) \cdot \mathsf{caus}'_s(\emptyset, P^\perp) \\ \overline{\mathsf{eff}_s}(M, P)(z_M, z_P) &:= \mathsf{eff}'_s(M_1, P_1) \cdot \mathsf{eff}'_s(M_2 P_2) \cdot \mathsf{eff}'_s(\emptyset, P^\perp) \,, \end{aligned} \tag{2.30}$$

where we have abused notation by equating each subset M_1 and M_2 of nodes with their induced subsystems of S via the state s.

This concludes all data necessary to define classical IIT. If the generalized definition of Section 2.8 is applied to this data, it yields precisely classical IIT 3.0 defined by Tononi et al. In Appendix 2.A, we explain in detail how our definition of IIT, equipped with this data, maps to the usual presentation of the theory.

2.10. Quantum IIT

In this section, we consider quantum IIT defined in (Zanardi et al., 2018). This is also a special case of the definition in terms of process theories we give in (Tull & Kleiner, 2021).

2.10.1. Systems

Similar to classical IIT, in quantum IIT systems are conceived as consisting of elements $\mathcal{H}_1, \ldots, \mathcal{H}_n$. Here, each element $\mathcal{H}_i$ is described by a finite dimensional Hilbert space and the state space of the system S is defined in terms of the element Hilbert spaces as

$$\mathsf{St}(S) = \mathcal{S}(\mathcal{H}_S) \qquad \text{with} \quad \mathcal{H}_S = \bigotimes_{i=1}^{n} \mathcal{H}_i \,,$$

where $\mathcal{S}(\mathcal{H}_S) \subset L(\mathcal{H}_S)$ describes the positive semidefinite Hermitian operators of unit trace on $\mathcal{H}_S$, i.e. density matrices. The time evolution of the system is again given by a time evolution operator, which here is assumed to be a trace preserving (and in (Zanardi et al., 2018) typically unital) completely positive map

$$\mathcal{T} : L(\mathcal{H}_S) \to L(\mathcal{H}_S) \,.$$

2.10.2. Subsystems, Decompositions and Cuts

Subsystems are again defined to consist of subsets M of the elements of the system, with corresponding Hilbert space $\mathcal{H}_M := \bigotimes_{i \in M} \mathcal{H}_i$. The time-evolution $\mathcal{T}_M : L(\mathcal{H}_M) \to L(\mathcal{H}_M)$ is defined as

$$\mathcal{T}_M(\rho) = \mathrm{tr}_{M^\perp} \left(\mathcal{T}(\mathrm{tr}_{M^\perp}(s) \otimes \rho) \right) ,$$

where $s \in \mathcal{S}(\mathcal{H}_S)$ and $\mathrm{tr}_{M^\perp}$ denotes the trace over the Hilbert space $\mathcal{H}_{M^\perp}$.

Decompositions are also defined via partitions $z = (D, D^\perp) \in \mathbb{D}_S$ of the set of elements N into two disjoint subsets D and $D^\perp$ whose union is N. For any such decomposition, the cut system $S^{(D,D^\perp)}$ is defined to have the same set of states but time evolution

$$\mathcal{T}^{(D,D^\perp)}(s) = \mathcal{T}\big(\mathrm{tr}_{D^\perp}(s) \otimes \fint_{D^\perp} \big) \,,$$

where $\fint_{D^\perp}$ is the maximally mixed state on $\mathcal{H}_{D^\perp}$, i.e. $\fint_{D^\perp} = \frac{1}{\dim(\mathcal{H}_{D^\perp})} 1_{\mathcal{H}_{D^\perp}}$.

2.10.3. Proto-Experiences

For any $\rho, \sigma \in \mathcal{S}(\mathcal{H}_S)$, the trace distance defined as

$$d(\rho, \sigma) = \tfrac{1}{2} \, \mathrm{tr}_S \left(\sqrt{(\rho - \sigma)^2} \right)$$

turns $(S(\mathcal{H}_S), d)$ into a metric space. The space of proto-experiences is defined based on this metric space as described in Example 2.3.2,

$$\mathbb{PE}(S) := \overline{S(\mathcal{H}_S)} \,.$$

2.10.4. Repertoires

We finally come to the definition of the cause- and effect repertoire. Unlike classical IIT, the definition in (Zanardi et al., 2018) does not consider virtual elements. Let a system S in state $s \in \mathsf{St}(S)$ be given. As in Section 2.9.5, we utilize maps caus'_s and eff'_s which here map subsystems M and P to $\mathsf{St}(P)$. They are defined as

$$\begin{aligned}\mathsf{eff}'_s(M,P) &= \mathrm{tr}_{P^\perp}\,\mathcal{T}\big(\,\mathrm{tr}_{M^\perp}(s)\otimes \doteq_{M^\perp}\big)\\ \mathsf{caus}'_s(M,P) &= \mathrm{tr}_{P^\perp}\,\mathcal{T}^\dagger\big(\,\mathrm{tr}_{M^\perp}(s)\otimes \doteq_{M^\perp}\big)\,,\end{aligned}$$

where $\mathcal{T}^\dagger$ is the Hermitian adjoint of $\mathcal{T}$. We then define

$$\begin{aligned}\mathsf{caus}_s(M,P) &:= \mathsf{caus}'_s(M,P)\otimes \mathsf{caus}'_s(\emptyset,P^\perp)\\ \mathsf{eff}(M,P) &:= \mathsf{eff}'_s(M,P)\otimes \mathsf{eff}'_s(\emptyset,P^\perp)\,,\end{aligned}$$

each with intensity 1, where $\emptyset$ again denotes the empty mechanism. Similarly, decompositions of these elements over $\mathbb{D}_{M,P}$ are defined as

$$\begin{aligned}\overline{\mathsf{caus}_s}(M,P)(z_M,z_P) &:= \mathsf{caus}'_s(M_1,P_1)\otimes \mathsf{caus}'_s(M_2,P_2)\otimes \mathsf{caus}'_s(\emptyset,P^\perp)\\ \overline{\mathsf{eff}_s}(M,P)(z_M,z_P) &:= \mathsf{eff}'_s(M_1,P_1)\otimes \mathsf{eff}'_s(M_2,P_2)\otimes \mathsf{eff}'_s(\emptyset,P^\perp)\,,\end{aligned}$$

again with intensity 1, where $z_M = (M_1,M_2)\in\mathbb{D}_M$ and $z_P=(P_1,P_2)\in\mathbb{D}_P$.

2.11. Extensions of Classical IIT

The physical systems to which IIT 3.0 may be applied are limited in a number of ways: they must have a discrete time-evolution, satisfy Markovian dynamics and exhibit a discrete set of states (A. B. Barrett & Mediano, 2019). Since many physical systems do not satisfy these requirements, if IIT is to be taken as a fundamental theory about reality, it must be extended to overcome these limitations.

In this section, we show how IIT can be redefined to cope with continuous time, non-Markovian dynamics and non-compact state spaces, by a redefinition of the maps (2.26) and (2.28) and, in the case of non-compact state spaces, a slightly different choice of (2.24), while leaving all of the remaining structure as it is. While we do not think that our particular definitions are satisfying as a general definition of IIT, these results show that the disentanglement of the essential mathematical structure of IIT from auxiliary tools (the particular definition of cause-effect repertoires used to date) can help to overcome fundamental mathematical or conceptual problems.

In Section 2.11.3, we also explain which solution to the problem of non-canonical metrics is suggested by our formalism.

2.11.1. Discrete Time and Markovian Dynamics

In order to avoid the requirement of a discrete time and Markovian dynamics, instead of working with the time evolution operator (2.18), we define the cause- and effect repertoires in reference to a given trajectory of a physical state $s \in \mathrm{St}(S)$. The resulting definitions can be applied independently of whether trajectories are being determined by Markovian dynamics in a particular application, or not.

Let $t \in \mathbb{I}$ denote the time parameter of a physical system. If time is discrete, $\mathbb{I}$ is an ordered set. If time is continuous, $\mathbb{I}$ is an interval of reals. For simplicity, we assume $0 \in \mathbb{I}$. In the deterministic case, a trajectory of a state $s \in \mathrm{St}(S)$ is simply a curve in $\mathrm{St}(S)$, which we denote by $(s(t))_{t\in\mathbb{I}}$ with $s(0) = s$. For probabilistic systems (such as neural networks with a probabilistic update rule), it is a curve of probability distributions $\mathcal{P}(S)$, which we denote by $(p(t))_{t\in\mathbb{I}}$, with $p(0)$ equal to the Dirac distribution δ_s. The latter case includes the former, again via Dirac distributions.

In what follows, we utilize the fact that in physics, state spaces are defined such that the dynamical laws of a system allow to determine the trajectory of each state. Thus for every $s \in \mathrm{St}(S)$, there is a trajectory $(p_s(t))_{t\in\mathbb{I}}$ which describes the time evolution of s.

The idea behind the following is to define, for every $M, P \in \mathrm{Sub}(S)$, a trajectory $p_s^{(P,M)}(t)$ in $\mathcal{P}(P)$ which quantifies how much the state of the purview P at time t is being constrained by imposing the state s at time $t = 0$ on the mechanism M. This gives an alternative definition of the maps (2.26) and (2.28), while the rest of classical IIT can be applied as before.

Let now $M, P \in \mathrm{Sub}(S)$ and $s \in \mathrm{St}(S)$ be given. We first consider the time evolution of the state $(s_M, v) \in \mathrm{St}(S)$, where s_M denotes the restriction of s to $\mathrm{St}(M)$ as before and where $v \in \mathrm{St}(M^\perp)$ is an arbitrary state of $M^\perp$. We denote the time evolution of this state by $p_{(s_M,v)}(t) \in \mathcal{P}(S)$. Marginalizing this distribution over $P^\perp$ gives a distribution on the states of P, which we denote as $p^P_{(s_M,v)}(t) \in \mathcal{P}(P)$. Finally, we average over v using the uniform distribution $\omega_{M^\perp}$. Because state spaces are finite in classical IIT, this averaging can be defined pointwise for every $w \in \mathrm{St}(P)$ by

$$p_s^{(P,M)}(t)(w) := \kappa \sum_{v\in\mathrm{St}(M^\perp)} p^P_{(s_M,v)}(t)(w)\, \omega_{M^\perp}(v)\,, \tag{2.31}$$

where κ is the unique normalization constant which ensures that $p_s^{(P,M)}(t) \in \mathcal{P}(P)$.

The probability distribution $p_s^{(P,M)}(t) \in \mathcal{P}(P)$ describes how much the state of the purview P at time t is being constrained by imposing the state s on M at time $t = 0$ as desired. Thus, for every $t \in \mathbb{I}$, we have obtained a mapping of two subsystems M, P to an element $p_s^{(P,M)}(t)$ of $\mathcal{P}(P)$ which has the same interpretation as the map (2.25) considered in classical IIT. If deemed necessary, virtual elements could be introduced just as in (2.26) and (2.28).

So far, our construction can be applied for any time $t \in T$. It remains to fix this freedom in the choice of time. For the discrete case, the obvious choice is to define (2.26)

and (2.28) in terms of neighbouring time-steps. For the continuous case, several choices exist. E.g., one could consider the positive and negative semi-derivatives of $p_s^{(P,M)}(t)$ at $t = 0$, in case they exist, or add an arbitrary but fixed time scale Δ to define the cause- and effect repertoires in terms of $p_s^{(P,M)}(t_0 \pm \Delta)$. However, the most reasonable choice is in our eyes to work with limits, in case they exist, by defining

$$\mathsf{eff}'_s(M, P) := \prod_{P_i \in P} \lim_{t \to \infty} p_s^{(P_i, M)}(t) \tag{2.32}$$

to replace (2.26) and

$$\mathsf{caus}'_s(M, P) := \kappa \prod_{M_i \in M} \lim_{t \to -\infty} p_s^{(P, M_i)}(t) \tag{2.33}$$

to replace (2.28). The remainder of the definitions of classical IIT can then be applied as before.

2.11.2. Discrete Set of States

The problem with applying the definitions of classical IIT to systems with continuous state spaces (e.g. neuron membrane potentials (A. B. Barrett & Mediano, 2019)) is that in certain cases, uniform probability distributions do not exist. E.g., if the state space of a system S consists of the positive real numbers $\mathbb{R}^+$, no uniform distribution can be defined which has a finite total volume, so that no uniform *probability* distribution ω_S exists.

It is important to note that this problem is less universal than one might think. E.g., if the state space of the system is a closed and bounded subset of $\mathbb{R}^+$, e.g. an interval $[a, b] \subset \mathbb{R}^+$, a uniform probability distribution can be defined using measure theory, which is in fact the natural mathematical language for probabilities and random variables. Nevertheless, the observation in (A. B. Barrett & Mediano, 2019) is correct that if a system has a non-compact continuous state space, ω_S might not exist, which can be considered a problem w.r.t. the above-mentioned working hypothesis.

This problem can be resolved for all well-understood physical systems by replacing the uniform probability distribution ω_S by some other mathematical entity which allows to define a notion of averaging states. An example is quantum theory (Section 2.10), whose state-spaces are continuous and non-compact. Here, the maximally mixed state $\frac{1}{=}_S$ plays the role of the uniform probability distribution. For all relevant classical systems with non-compact state spaces (whether continuous or not), the same is true: There exists a canonical uniform measure μ_S which allows to define the cause-effect repertoires similar to the last section, as we now explain. Examples for this canonical uniform measure are the Lebesgue measure for subsets of $\mathbb{R}^n$ (Rudin, 2006), or the Haar measure for locally compact topological groups (Salamon, 2016) such as Lie-groups.

In what follows, we explain how the construction of the last section needs to be modified in order to be applied to this case. In all relevant classical physical theories, $\mathsf{St}(S)$

is a metric space in which every probability measure is a Radon measure, in particular locally finite, and where a canonical locally finite uniform measure μ_S exists. We define $\mathcal{P}_1(S)$ to be the space of probability measures whose first moment is finite. For these, the first Wasserstein metric (or 'Earth Mover's Distance') W_1 exists, so tat $(\mathcal{P}_1(S), W_1)$ is a metric space.

As before, the dynamical laws of the physical systems determine for every state $s \in \mathrm{St}(S)$ a time evolution $p_s(t)$, which here is an element of $\mathcal{P}_1(S)$. Integration of this probability measure over $\mathrm{St}(P^\perp)$ yields the marginal probability measure $p_s^P(t)$. As in the last section, we may consider these probability measures for the state $(s_M, v) \in \mathrm{St}(S)$, where $v \in \mathrm{St}(M^\perp)$. Since μ_S is not normalizable, we cannot define $p_s^{(P,M)}(t)$ as in (2.31), for the result might be infinite.

Using the fact that μ_S is locally finite, we may, however, define a somewhat weaker equivalent. To this end, we note that for every state $s_{M^\perp}$, the local finiteness of $\mu_{M^\perp}$ implies that there is a neighbourhood $N_{s,M^\perp}$ in $\mathrm{St}(M^\perp)$ for which $\mu_{M^\perp}(N_{s,M^\perp})$ is finite. We choose a sufficiently large neighbourhood which satisfies this condition. Assuming $p^P_{(s_M,v)}(t)$ to be a measurable function in v, for every A in the σ-algebra of $\mathrm{St}(M^\perp)$, we can thus define

$$p_s^{(P,M)}(t)(A) := \kappa \int_{N_{s,M^\perp}} p^P_{(s_M,v)}(t)(A)\, d\mu_{M^\perp}(v)\,, \tag{2.34}$$

which is a finite quantity. The $p_s^{(P,M)}(t)$ so defined is non-negative, vanishes for $A = \emptyset$ and satisfies countable additivity. Hence it is a measure on $\mathrm{St}(P)$ as desired, but might not be normalizable.

All that remains for this to give a cause-effect repertoire as in the last section, is to make sure that any measure (normalized or not) is an element of $\mathbb{PE}(S)$. The theory is flexible enough to do this by setting $d(\mu,\nu) = |\mu - \nu|(\mathrm{St}(P))$ if either μ or ν is not in $\mathcal{P}_1(S)$, and $W_1(\mu,\nu)$ otherwise. Here, $|\mu - \nu|$ denotes the total variation of the signed measure $\mu - \nu$, and $|\mu - \nu|(\mathrm{St}(P))$ is the volume thereof (Encyclopedia of Mathematics, 2013; Halmos, 1974). While not a metric space any more, the tuple $(\mathcal{M}(S), d)$, with $\mathcal{M}(S)$ denoting all measures on $\mathrm{St}(S)$, can still be turned into a space of proto-experiences as explained in Example 2.3.2. This gives

$$\mathbb{PE}(S) := \overline{\mathcal{M}(S)}$$

and finally allows to construct cause-effect repertoires as in the last section.

2.11.3. Non-Canonical Metrics

Another criticism of IIT's mathematical structure mentioned (A. B. Barrett & Mediano, 2019) is that the metrics used in IIT's algorithm are, to a certain extend, chosen arbitrarily. Different choices indeed imply different results of the algorithm, both concerning the quantity and quality of experience, which can be considered problematic.

The resolution of this problem is, however, not so much a technical as a conceptual or philosophical task, for what is needed to resolve this issue is a justification of why a particular metric should be used. Various justifications are conceivable, e.g. identification of desired behaviour of the algorithm when applied to simple systems. When considering our mathematical reconstruction of the theory, the following natural justification offers itself.

Implicit in our definition of the theory as a map from systems to experience spaces is the idea that the mathematical structure of experiences spaces (Definition 2.3.1) reflects the phenomenological structure of experience. This is so, most crucially, for the distance function d, which describes how similar two elements of experience spaces are. Since every element of an experience space corresponds to a conscious experience, it is naturally to demand that the similarly of the two mathematical objects should reflect the similarity of the experiences they describe. Put differently, the distance function d of an experience space should in fact mirror (or "model") the similarity of conscious experiences as experienced by an experiencing subject.

This suggests that the metrics d used in the IIT algorithm should, ultimately, be defined in terms of the phenomenological structure of similarity of conscious experiences. For the case of colour qualia, this is in fact feasible (Kleiner, 2020b, Example 3.18), (R. Kuehni, 2010; Sharma, Wu, & Dalal, 2004). In general, the mathematical structure of experience spaces should be intimately tied to the phenomenology of experience, in our eyes.

2.12. Summary & Outlook

In this article, we have propounded the mathematical structure of Integrated Information Theory. First, we have studied which exact structures the IIT algorithm uses in the mathematical description of physical systems, on the one hand, and in the mathematical description of conscious experience, on the other. Our findings are the basis of definitions of a physical system class $\mathbf{Sys}$ and a class $\mathbf{Exp}$ of experience spaces, and allowed us to view IIT as a map $\mathbf{Sys} \to \mathbf{Exp}$.

Next, we needed to disentangle the essential mathematics of the theory from auxiliary formal tools used in the contemporary definition. To this end, we have introduced the precise notion of decomposition of elements of an experience space required by the IIT algorithm. The pivotal cause-effect repertoires are examples of decompositions so defined, which allowed us to view any particular choice, e.g. the one of 'classical' IIT developed by Tononi et. al., or the one of 'quantum' IIT recently introduced by Zanardi et. al. as data provided to a general IIT algorithm.

The formalization of cause-effect repertoires in terms of decompositions then led us to define the essential ingredients of IIT's algorithm concisely in terms of integration levels, integration scalings and cores. These definitions describe and unify recurrent mathematical operations in the contemporary presentation, and finally allowed to define IIT completely in terms of a few lines of definition.

Throughout the paper, we have taken great care to make sure our definitions repro-

duce exactly the contemporary version of IIT 3.0. The result of our work is a mathematically rigorous and general definition of Integrated Information Theory. This definition can be applied to any meaningful notion of systems and cause-effect repertoires, and we have shown that this allows one to overcome most of the mathematical problems of the contemporary definition identified to date in the literature.

We believe that our mathematical reconstruction of the theory can be the basis for refined mathematical and philosophical analysis of IIT. We also hope that this mathematisation may make the theory more amenable to study by mathematicians, physicists, computer scientists and other researchers with a strongly formal background.

2.12.1. Process Theories

Our generalization of IIT is axiomatic in the sense that we have only included those formal structures in the definition which are necessary for the IIT algorithm to be applied. This ensured that our reconstruction is as general as possible, while still true to IIT 3.0. As a result, several notions used in classical IIT, e.g., system decomposition, subsystems or causation, are merely defined abstractly at first, without any reference to the usual interpretation of these concepts in physics.

In the related article (Tull & Kleiner, 2021), we show that these concepts can be meaningfully defined in any suitable *process theory* of physics, formulated in the language of *symmetric monoidal categories*. This approach can describe both classical and quantum IIT and yields a complete formulation of contemporary IIT in a categorical framework.

2.12.2. Further Development of IIT

IIT is constantly under development, with new and refined definitions being added every few years. We hope that our mathematical analysis of the theory might help to contribute to this development. E.g., the working hypothesis that IIT is a fundamental theory, i.e. describes reality as it is, implies that technical problems of the theory need to be resolved. We have shown that our formalization allows one to address the technical problems mentioned in the literature. However, there are others which we have not addressed in this paper.

Most crucially, the IIT algorithm uses a series of maximalization and minimalization operations, unified in the notion of *core* subsystems in our formalization. In general, there is no guarantee that these operations lead to unique results, neither in classical nor quantum IIT. Using different cores has major impact on the output of the algorithm, including the Φ value, which is a case of ill-definedness.[3]

Furthermore, the contemporary definition of IIT as well as our formalization rely on there being a finite number of subsystems of each system, which might not be the case

[3] The problem of 'unique existence' has been studied extensively in category theory using *universal properties* and the notion of a *limit*. Rather than requiring that each $E \in \mathbb{E}$ come with a metric, it may be possible to alter the IIT algorithm into a well-defined categorical form involving limits to resolve this problem.

in reality. Our formalisation may be extendable to the infinite case by assuming that every system has a fixed but potentially infinite indexing set $\mathrm{Sub}(S)$, so that each $\mathrm{Sub}_s(S)$ is the image of a mapping $\mathrm{Sub}(S) \times \mathsf{St}(S) \to \mathbf{Sys}$, but we have not considered this in detail in this paper.

Finally, concerning more operational questions, it would be desirable to develop the connection to empirical measures such as the Perturbational Complexity Index PCI (Casa rotto et al., 2016; Casali et al., 2013) in more detail, as well as to define a controlled approximation of the theory whose calculation is less expensive. Both of these tasks may be achievable by substituting parts of our formalization with simpler mathematical structure.

On the conceptual side of things, it would be desirable to have a more proper understanding of how the mathematical structure of experiences spaces corresponds to the phenomenology of experience, both for the general definition used in our formalization – which comprises the minimal mathematical structure which is required for the IIT algorithm to be applied – and the specific definitions used in classical and quantum IIT. In particular, it would be desirable to understand how it relates to the important notion of qualia, which is often asserted to have characteristic features such as ineffability, intrinsicality, non-contextuality, transparency or homogeneity (Metzinger, 2006). For a first analysis towards this goal, cf. (Kleiner, 2020b). A first proposal to add additional structure that accounts for relations between elements of consciousness in the case of spatial experiences was recently given in (A. Haun & Tononi, 2019).

We would like to thank the organizers and participants of the *Workshop on Information Theory and Consciousness* at the Centre for Mathematical Sciences of the University of Cambridge, of the *Modelling Consciousness Workshop* in Dorfgastein and of the *Models of Consciousness Conference* at the Mathematical Institute of the University of Oxford for discussions on this topic. Much of this work was carried out while Sean Tull was under the support of an EPSRC Doctoral Prize at the University of Oxford, from November 2018 to July 2019, and while Johannes Kleiner was under the support of postdoctoral funding at the Institute for Theoretical Physics of the Leibniz University of Hanover. We would like to thank both institutions.

Appendix

2.A. Comparison with Standard Presentation of IIT 3.0

In Section 2.9, we have defined the system class and cause-effect repertoires which underlie classical IIT. The goal of this appendix is to explain in detail why applying our definition of the IIT algorithm yields IIT 3.0 defined by Tononi et al. In doing so, we will mainly refer to the terminology used in (Tononi, 2015), (Mayner et al., 2018), (Oizumi et al., 2014) and (Marshall et al., 2016). We remark that a particularly detailed presentation of the algorithm of the theory, and of how the cause and effect repertoire are calculated, is given in the supplementary material S1 of (Mayner et al., 2018).

2.A.1. Physical Systems

The systems of classical IIT are given in Section 2.9.1. They are often represented as graphs whose nodes are the elements $S_1, \ldots, S_n$ and edges represent functional dependence, thus describing the time evolution of the system as a whole, which we have taken as primitive in (2.18). This is similar to the presentation of the theory in terms of a transition probability function

$$p : \mathsf{St}(S) \times \mathsf{St}(S) \to [0, 1]$$

in (Marshall et al., 2016). For each probability distribution $\tilde{p}$ over $\mathsf{St}(S)$, this relates to our time evolution operator T via

$$T(\tilde{p})[v] := \sum_{w \in \mathsf{St}(S)} p(v, w)\, \tilde{p}(w)\ .$$

2.A.2. Cause-Effect Repertoires

In contemporary presentations of the theory ((Marshall et al., 2016, p. 14) or (Tononi, 2015)), the effect repertoire is defined as

$$p_{\text{effect}}(z_i, m_t) := \frac{1}{|\Omega_{M^c}|} \sum_{m^c \in \Omega_{M^c}} p\big(\, z_i \,|\, \mathrm{do}(m_t, m^c)\,\big) \qquad z_i \in \Omega_{Z_i} \tag{2.35}$$

and

$$p_{\text{effect}}(z, m_t) := \prod_{i=1}^{|z|} p_{\text{effect}}(z_i, m_t)\ . \tag{2.36}$$

Here, m_t denotes a state of the mechanism M at time t. M^c denotes the complement of the mechanism, denoted in our case as $M^\perp$, Ω_{M^c} denotes the state space of the complement, and m^c an element thereof. Z_i denotes an element of the purview Z (designated by P in our case), Ω_{Z_i} denotes the state space of this element, z_i a state of this element and z a state of the whole purview. $|\Omega_{M^c}|$ denotes the cardinality of the state space of M^c, and $|z|$ equals the number of elements in the purview. Finally, the expression $\mathrm{do}(m_t, m^c)$ denotes a variant of the so-called "do-operator". It indicates that the state of the system, here at time t, is to be set to the term in brackets. This is called *perturbing the system* into the state (m_t, m^c). The notation $p(z_i|\mathrm{do}(m_t, m^c))$ then gives the probability of finding the purview element in the state z_i at time $t+1$ given that the system is prepared in the state (m_t, m^c) at time t.

In our notation, the right hand side of (2.35) is exactly given by the right-hand side of (2.25), i.e. $\mathsf{eff}'_s(M, P_i)$. The system is prepared in a uniform distribution on M^c (described by the sum and prefactor in (2.35)) and with the restriction s_M of the system state, here denoted by m_t, on M. Subsequently, T is applied to evolve the system to time $t+1$, and the marginalization $\langle P_i^\perp|$ throws away all parts of the states except those of the purview element P_i (denoted above as Z_i). In total, (2.25) is a probability distribution on the states of the purview element. When evaluating this probability distribution at one particular state z_i of the element, one obtains the same numerical value as (2.35). Finally, taking the product in (2.36) corresponds exactly to taking the product in (2.26).

Similarly, the cause repertoire is defined as ((Marshall et al., 2016, p. 14) or (Tononi, 2015))

$$p_{\mathrm{cause}}(z|m_{i,t}) := \frac{\sum_{z^c \in \Omega_{Z^c}} p\big(\, m_{i,t} \,|\, \mathrm{do}(z, z^c)\,\big)}{\sum_{s \in \Omega_S} p\big(\, m_{i,t} \,|\, \mathrm{do}(s)\,\big)} \qquad z \in \Omega_{Z_{t-1}} \tag{2.37}$$

and

$$p_{\mathrm{cause}}(z|m_t) := \frac{1}{K} \prod_{i=1}^{|m_t|} p_{\mathrm{cause}}(z|m_{i,t})\,, \tag{2.38}$$

where m_i denotes the state of one element of the mechanism M, with the subscript t indicating that the state is considered at time t. Z again denotes a purview, z is a state of the purview and $\Omega_{Z_{t-1}}$ denotes the state space of the purview, where the subscript indicates that the state is considered at time $t-1$. K denotes a normalization constant and $|m_t|$ gives the number of elements in M.

Here, the whole right hand side of (2.37) gives the probability of finding the purview in state z at time $t-1$ if the system is prepared in state $m_{i,t}$ at time t. In our terminology this same distribution is given by (2.27), where λ is the denominator in (2.37). Taking the product of these distributions and re-normalising is then precisely (2.28).

As a result, the cause and effect repertoire in the sense of (Oizumi et al., 2014) correspond precisely in our notation to $\mathsf{caus}'_s(M, P)$ and $\mathsf{eff}'_s(M, P)$, each being distributions over $\mathrm{St}(P)$. In (Mayner et al., 2018, S1), it is explained that these need to be extended by the unconstrained repertoires before being used in the IIT algorithm, which in our formalization is done in (2.29), so that the cause-effect repertoires are now distributions over

$\mathrm{St}(S)$. These are in fact precisely what are called the *extended* cause and effect repertoires or *expansion to full state space* of the repertoires in (Oizumi et al., 2014).

The behaviour of the cause- and effect-repertoires when decomposing a system is described, in our formalism, by decompositions (Definition 2.4.1). Hence a decomposition $z \in \mathbb{D}_S$ is what is called a *parition* in the classical formalism. For the case of classical IIT, a decomposition is given precisely by a partition of the set of elements of a system, and the cause-effect repertoires belonging to the decomposition are defined in (2.30), which corresponds exactly to the definition

$$p^{\mathrm{cut}}_{\mathrm{cause}}(z|m_t) = p_{\mathrm{cause}}(z^{(1)}|m_t^{(1)}) \times p_{\mathrm{cause}}(z^{(2)}|m_t^{(2)})$$

in (Marshall et al., 2016), when expanded to the full state space, and equally so for the effect repertoire.

2.A.3. Algorithm - Mechanism Level

Next, we explicitly unpack our form of the IIT algorithm to see how it compares in the case of classical IIT with (Oizumi et al., 2014). In our formalism, the integrated information φ of a mechanism M of system S when in state s is

$$\varphi^{\max}(M) = \| \, \mathbb{C}_{S,s}(M) \, \| \tag{2.39}$$

defined in Equation (2.10). This definition conjoins several steps in the definition of classical IIT. To explain why it corresponds exactly to classical IIT, we disentangle this definition step by step.

First, consider $\mathsf{caus}_s(M, P)$ in Equation (2.9). This is, by definition, a decomposition map. The calculation of the integration level of this decomposition map, cf. Equation (2.5), amounts to comparing $\mathsf{caus}_s(M, P)$ to the cause-effect repertoire associated with every decomposition using the metric of the target space $\mathbb{PE}(S)$, which for classical IIT is defined in (2.24) and Example 2.3.2, so that the metric d used for comparison is indeed the Earth Mover's Distance. Since cause-effect repertoires have, by definition, unit intensity, the factor r in the definition (2.1) of the metric does not play a role at this stage. Therefore, the integration level of $\mathsf{caus}_s(M, P)$ is exactly the *integrated cause information*, denoted as

$$\varphi^{\mathrm{MIP}}_{\mathrm{cause}}(y_t, Z_{t-1})$$

in (Tononi, 2015), where y_t denotes the (induced state of the) mechanism M in this notation, and Z_{t-1} denotes the purview P. Similarly, the integration level of $\mathsf{eff}_s(M, P)$ is exactly the *integrated effect information*, denoted as

$$\varphi^{\mathrm{MIP}}_{\mathrm{effect}}(y_t, Z_{t+1}) \, .$$

The integration scaling in (2.10) simply changes the intensity of an element of $\mathbb{PE}(S)$ to match the integration level, using the scalar multiplication, which is important for the system level definitions. When applied to $\mathsf{caus}_s(M, P)$, this would result in an element of $\mathbb{PE}(S)$ whose intensity is precisely $\varphi^{\mathrm{MIP}}_{\mathrm{cause}}(y_t, Z_{t-1})$.

Consider now the collections (2.9) of decomposition maps. Applying Definition 2.5.2, the core of $\mathsf{caus}_s(M)$ is that purview P which gives the decomposition $\mathsf{caus}_s(M,P)$ with the highest integration level, i.e. with the highest $\varphi^{\mathrm{MIP}}_{\mathrm{cause}}(y_t, Z_{t-1})$. This is called the *core cause* P^c of M, and similarly the core of $\mathsf{eff}_s(M)$ is called the *core effect* P^e of M.

Finally, to fully account for (2.10), we note that the integration scaling of a pair of decomposition maps rescales both elements to the minimum of the two integration levels. Hence the integration scaling of the pair $(\mathsf{caus}_s(M,P), \mathsf{eff}(M,P'))$ fixes the scalar value of both elements to be exactly the *integrated information*, denoted as

$$\varphi(y_t, Z_{t\pm1}) = \min\left(\varphi^{\mathrm{MIP}}_{\mathrm{cause}}, \varphi^{\mathrm{MIP}}_{\mathrm{effect}}\right)$$

in (Tononi, 2015), where $P = Z_{t+1}$ and $P' = Z_{t-1}$.

In summary, the following operations are combined in Equation (2.10). The core of $(\mathsf{caus}_s(M), \mathsf{eff}_s(M))$ picks out the core cause P^c and core effect P^e. The core integration scaling subsequently considers the pair $(\mathsf{caus}_s(M,P^c), \mathsf{eff}(M,P^e))$, called *maximally irreducible cause-effect repertoire*, and determines the integration level of each by analysing the behaviour with respect to decompositions. Finally, it rescales both to the minimum of the integration levels. Thus it gives exactly what is called $\varphi^{\max}$ in (Tononi, 2015). Using, finally, the definition of the intensity of the product $\mathbb{PE}(S) \times \mathbb{PE}(S)$ in Definition 2.3.3, this implies (2.39). The concept of M in our formalization is given by the tuple

$$\mathbb{C}_{S,s}(M) := \left((\mathsf{caus}_s(M,P^c), \varphi^{\max}(M)), (\mathsf{eff}_s(M,P^e), \varphi^{\max}(M))\right)$$

i.e. the pair of maximally irreducible repertoires scaled by $\varphi^{\max}(M)$. This is equivalent to what is called a *concept*, or sometimes *quale sensu stricto*, in classcial IIT (Tononi, 2015), and denoted as $q(y_t)$.

We finally remark that it is also possible in classical IIT that a cause repertoire value $\mathsf{caus}_s(M,P)$ vanishes (Remark 2.9.1). In our formalization, it would hence be represented by $(\omega_S, 0)$ in $\mathbb{PE}(S)$, so that $d(\mathsf{caus}_s(M,P), q) = 0$ for all $q \in \mathbb{E}(S)$ according to (2.1), which certainly ensures that $\varphi^{\mathrm{MIP}}_{\mathrm{cause}}(M,P) = 0$.

2.A.4. Algorithm - System Level

We finally explain how the system level definitions correspond to the usual definition of classical IIT.

The Q-shape $\mathbb{Q}_s(S)$ is the collection of all concepts specified by the mechanisms of a system. Since each concept has intensity given by the corresponding integrated information of the mechanism, this makes $\mathbb{Q}_s(S)$ what is usually called the *conceptual structure* or *cause-effect structure*. In (Oizumi et al., 2014), one does not include a concept for any mechanism M with $\varphi^{\max}(M) = 0$. This manual exclusion is unnecessary in our case because the mathematical structure of experience spaces implies that mechanisms with $\varphi^{\max}(M) = 0$ should be interpreted as having no conscious experience, and the algorithm in fact implies that they have 'no effect'. Indeed we will now see that they do not contribute to the distances in $\mathbb{E}(S)$ or any Φ values, and so we do not manually exclude them.

When comparing $\mathbb{Q}_s(S)$ with the Q-shape (2.13) obtained after replacing S by any of its cuts, it is important to note that both are elements of $\mathbb{E}(S)$ defined in (2.12), which is a product of experience spaces. According to Definition 2.3.3, the distance function on this product is

$$d(\mathbb{Q}_s(S), \mathbb{Q}_s(S^z)) := \sum_{M \in \mathrm{Sub}(S)} d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M)) \,.$$

Using Definition 2.3.2 and the fact that each concept's intensity is $\varphi^{\max}(M)$ according to the mechanism level definitions, each distance $d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M))$ is equal to

$$\begin{aligned} \varphi^{\max}(M) \cdot \big(& d\big(\mathsf{caus}_s(M, P^c_M), \mathsf{caus}^z_s(M, P^{z,c}_M)\big) \\ & + d\big(\mathsf{eff}_s(M, P^e_M), \mathsf{eff}^z_s(M, P^{z,e}_M)\big)\big) \,, \end{aligned} \tag{2.40}$$

where $\varphi^{\max}(M)$ denotes the integrated information of the concept in the original system S, and where the right-hand cause and effect repertoires are those of S^z at its own core causes and effects for M. The factor $\varphi^{\max}(M)$ ensures that the distance used here corresponds precisely to the distance used in (Oizumi et al., 2014), there called the *extended Earth Mover's Distance*. If the integrated information $\varphi^{\max}(M)$ of a mechanism is non-zero, it follows that $d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M)) = 0$ as mentioned above, so that this concept does not contribute.

We remark that in (Mayner et al., 2018, S1), an additional step is mentioned which is not described in any of the other papers we consider. Namely, if the integrated information of a mechanism is non-zero before cutting but zero after cutting, what is compared is not the distance of the corresponding concepts as in (2.40), but in fact the distance of the original concept with a special null concept, defined to be the unconstrained repertoire of the cut system. We have not included this step in our definitions, but it could be included by adding a choice of distinguished point to Example 2.3.2 and redefining the metric correspondingly.

In Equation (2.14) the above comparison is being conducted for every subsystem of a system S. The subsystems of S are what is called *candidate systems* in (Oizumi et al., 2014), and which describe that 'part' of the system that is going to be conscious according to the theory (cf. below). Crucially, candidate systems are subsystems of S, whose time evolution is defined in (2.22). This definition ensures that the state of the elements of S which are not part of the candidate system are fixed in their current state, i.e. constitute *background conditions* as required in the contemporary version of classcial IIT (Mayner et al., 2018).

Equation (2.14) then compares the Q-shape of every candidate system to the Q-shape of all of its cuts, using the distance function described above, where the cuts are defined in (2.23). The cut system with the smallest distance gives the system-level *minimum information partition* and the *integrated (conceptual) information* of that candidate system, denoted as $\Phi(x_t)$ in (Tononi, 2015).

The core integration scaling finally picks out that candidate system with the largest integrated information value. This candidate system is the *major complex* M of S, the

part of S which is conscious according to the theory as part of the *exclusion postulate* of IIT. Its Q-shape is the *maximally irreducible conceptual structure (MICS)*, also called *quale sensu lato*. The overall *integrated conceptual information* is, finally, simply the intensity of $\mathbb{E}(S, s)$ as defined in (2.14),

$$\Phi(S, s) = \mathbb{E}(S, s) \, .$$

2.A.5. Constellation in Qualia Space

Expanding our definitions, and denoting the major complex by M with state $m = s|_M$, in our terminology the actual experience of the system S state s is

$$\mathbb{E}(S, s) := \frac{\Phi(M, m)}{\| \, \mathbb{Q}_m(M) \, \|} \cdot \mathbb{Q}_m(M) \, . \tag{2.41}$$

This encodes the Q-shape $\mathbb{Q}_m(M)$, i.e. the maximally irreducible conceptual structure of the major complex, sometimes called *quale sensu lato*, which is taken to describe the quality of conscious experience. By construction it also encodes the integrated conceptual information of the major complex, which captures its intensity, since we have $\| \, \mathbb{E}(S, s) \, \| = \Phi(M, m)$. The rescaling of $\mathbb{Q}_m(M)$ in (2.41) leaves the relative intensities of the concepts in the MICS intact. Thus $\mathbb{E}(S, s)$ is the *constellation of concepts in qualia space* $\mathbb{E}(M)$ of (Oizumi et al., 2014).

3. Integrated Information in Process Theories

Sean Tull, Johannes Kleiner[1]

3.1. Introduction

Integrated Information Theory (IIT) is a theory of consciousness proposed and developed by Giulio Tononi and collaborators (Tononi, 2008; Oizumi et al., 2014). Originally defined in terms of a numerical measure Φ representing the level of phenomenal consciousness of a system (Tononi, 2004; P. A. Mediano, Seth, & Barrett, 2019), the most recent version of the theory, IIT 3.0, now employs an algorithm which claims to determine in addition which part of a system is conscious, and what it is conscious of.

In this article we show how the key concepts of IIT, including systems, integration and causation, can be studied naturally in the language of physical *process theories*, which are mathematically described as *symmetric monoidal categories*. Process theories come with an intuitive but rigorous graphical calculus (Selinger, 2011) which allows us to present many aspects of IIT in a simple pictorial fashion.

The constructions we provide in this article can be applied to any suitable process theory to yield a notion of *generalised IIT* as defined by the authors in a companion article (Kleiner & Tull, 2021). This allows us to extend IIT to new physical settings. As special cases, choosing the process theory of classical probabilistic processes essentially yields the usual IIT 3.0 in the sense of (Oizumi et al., 2014). Starting instead from the theory of quantum processes gives the *Quantum Integrated Information Theory* defined

[1]Published as: Tull, S., & Kleiner, J. (2021). Integrated Information in Process Theories: Towards Categorical IIT. *Journal of Cognitive Science*, 22, 2, 92–123. (Tull & Kleiner, 2021)

by Zanardi, Tomka and Venuti (Zanardi et al., 2018), which was another motivation for this work.

Independently of consciousness itself, our constructions provide a possible foundation for a general theory of integrated or 'holistic' behaviour within process theories, i.e. monoidal categories, which may be of interest to a broad range of fields. For example, neural net-like systems that achieve a task using a high degree of integration should be more efficient than fully modular ones, in that they require fewer neurons for the same task, and indeed integrated behaviour has been shown to evolve in simple models of biological organisms (Albantakis, Hintze, Koch, Adami, & Tononi, 2014). The methods of IIT have been applied generally in the study of integration in information processing systems, including treatments of autonomy (Marshall, Kim, Walker, Tononi, & Albantakis, 2017), causation (Albantakis, Marshall, Hoel, & Tononi, 2017), and state differentiation (Marshall et al., 2016).

3.1.1. Background: Mathematical Consciousness Science

The background for our work is in the growing field of Mathematical Consciousness Science (MCS), which aims to apply formal and mathematical tools in order to resolve open problems in the scientific study of consciousness. One major goal thereby is to expose and improve the mathematical structure of neuroscientific theories of consciousness so as to allow quantifiable comparison between competing models, generate novel experimental predictions, and to provide a thorough foundation for further development and combination of theories. More foundationally, it aims to uncover how consciousness relates to the physical world in terms of empirically grounded and philosophically motivated scientific theories. Progress in this direction is essential for resolving medical challenges (most notably, improving the understanding of neurological, psychiatric and psychological disorders (Michel et al., 2019)) and ethical reasons (for example the detection of consciousness in anesthetized or non-communicating patients (Alkire, Hudetz, & Tononi, 2008; Fink, Wiese, & Windt, 2018)), and could generate new advances in AI (artificial implementation of consciousness-related functions, for example (McDermott, 2007)).

A crucial cornerstone in this program is the representation of conscious experience in terms of a mathematical spaces, and to expound theories of consciousness as mappings from a mathematical description of physical systems to these spaces. Early precursors of the former are quality spaces (Beals, Krantz, & Tversky, 1968; A. Clark, 1993, 2000) which make use of just noticeable difference between stimuli to construct a representation of mental qualities and similarities between them. In the companion article (Kleiner & Tull, 2021), we provide a definition of an *experience space* that builds upon quality spaces while being geared at precisely what is required to flash out IIT as a mathematical mapping of the just-mentioned kind.

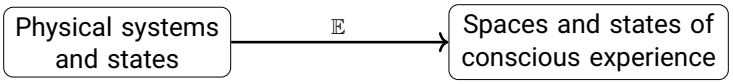
This contributes to the exploration and application of category theory as a framework for theories of consciousness (Tsuchiya, Taguchi, & Saigo, 2016; Northoff et al., 2019; Ehresmann, 2012). Category theory itself provides a natural language for describing

mappings between scientific domains, such as domains of physical systems and those modelling phenomenal experiences. Its emphasis on processes between systems in particular makes it ideal for describing theories and experimental findings which relate consciousness to dynamical processes, as discussed for example in (Fekete & Edelman, 2011; Wiese & Friston, 2021; Grindrod, 2018). The use of monoidal categories in this article additionally allows us to treat compositional aspects of systems and processes, which are central to theories such as IIT.

3.1.2. A primer on Integrated Information Theory

Though the majority of the article is self-contained and requires no prior knowledge of the theory, for context we include here a short introduction to IIT 3.0 (Oizumi et al., 2014), as formulated in its general form in our companion article (Kleiner & Tull, 2021) to which we refer for a more detailed presentation of the theory.

Any generalised IIT, including IIT 3.0, takes as input a given class of physical systems S, each with a given state space $\mathsf{St}(S)$, and specifies a map $\mathbb{E}$ which provides each system with a space describing its possible conscious experiences. Additionally, for each state $s \in \mathsf{St}(S)$ the theory specifies a particular experience $\mathbb{E}(s) \in \mathbb{E}(S)$ which the system will have in that state:



In IIT 3.0 the nature of this mapping derives from a number of essential properties–so called 'axioms'–which are postulated to characterize every conscious experience. Next to integration and information, these axioms include intrinsic existence, composition and exclusion (Tononi, 2015). These axioms are being translated into formal requirements. To this end, comparably simple physical systems are considered. These consist of a set of elements (or 'nodes'), each usually with only two states (on or off), and come with a discrete Markovian time evolution which is often described via a given causal graph. The prototypical example would be a human brain, in which the nodes represent neurons and their firing. The result of the translation process is the algorithm of IIT 3.0, i.e. the map $\mathbb{E}$ when applied to classical physical systems.

Starting from such a system S along with its current state s, the theory then specifies a set of probability distributions known as the *cause-effect repertoire*. For each pair of subsystems M, P ('mechanism' and 'purview') of S, the cause repertoire $\mathsf{caus}(M, P)$ is a distribution specifying how the current state of M constrains the state of P in the previous time-step, and similarly the effect repertoire $\mathsf{eff}(M, P)$ addresses the next time-step instead.

In the IIT algorithm one goes on how to calculate how 'integrated' each of these repertoires are by comparing them against repertoires obtained instead by 'cutting' the (evolution of the) system into various parts, by removing causal connections between them. For each mechanism M one determines which purviews give the most integrated values of $\mathsf{caus}(M, P)$ and $\mathsf{eff}(M, P)$, and these repertoire values (along with their level of

integration) determine a *concept* for that mechanism. The weighted collection of these concepts determines the entity $\mathbb{E}(s)$, also known as the *Q-shape* of the system, which is claimed to specify its total conscious experience. In particular this Q-Shape comes with its own level of integration, denoted $\Phi(s)$, which describes 'how conscious' the system is as a whole. A final 'exclusion' step enforces that only the subsystem of S with the highest Φ value will in fact be conscious.

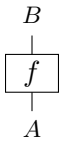
In the article (Kleiner & Tull, 2021) we show how to define a broad class of generalisations of IIT, in which for example the repertoires need no longer be described by probability distributions, but the states of a general physical theory. In the present article we describe how such IITs may be defined starting from any physical process theory. To do so we define the key notions of any IIT within such a setting, namely causal relations and their integration.

3.1.3. Structure of article

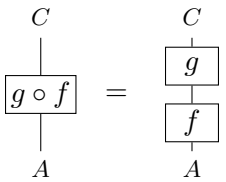
The article is structured as follows. We introduce process theories in Section 3.2 and then use them to describe the key notions from IIT – decompositions of objects (Section 3.3), systems (Section 3.4) and cause and effect repertoires (Section 3.5). We summarise how to define a generalised IIT from a process theory in Section 3.6 before giving examples in Section 3.7 and discussing future work in Section 3.8. The appendix contains some initial steps in developing a general study of integration in monoidal categories.

3.2. Process Theories

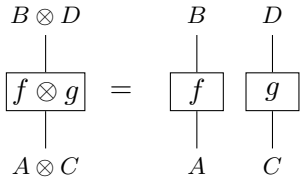
We begin by introducing the framework of *process theories* used throughout this work; for more detailed introductions we refer to (Coecke & Paquette, 2010; Coecke & Kissinger, 2017). The basic ingredients of such a theory are *objects* and *processes* between them. We depict a process from the object A to the object B as a box:



These processes may be *composed* together to form new ones in several ways. Firstly, given a process such as f above, and any other process g from B to C, we may compose them 'in sequence' to form a new one from A to C, denoted:

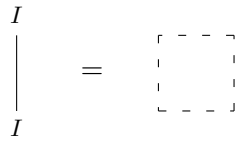


Secondly, we may compose processes in parallel. Any two objects A, B may be combined into a single object $A \otimes B$. Moreover any processes f from A to B, and g from C to D may be placed 'side-by-side' to form a new process:



from $A \otimes C$ to $B \otimes D$. More generally, by combining these operations, many processes may all be plugged together to form more complex diagrams describing a single composite process.

As a convenience, any process theory is taken to come with the following. Firstly, any object A come with an *identity process*, depicted as a blank wire on A, which 'does nothing' in that composing with it via $\circ$ leaves any process as it is. Secondly, it has a *trivial object*, denoted I, which leaves objects alone when combining under $\otimes$. We depict I as empty space:

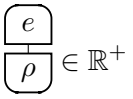


Finally, we formally assume the presence of a special process $\times$ which allows us to 'swap' any pair of wires over each other, along with a set of rules saying roughly that diagrams in the above sense are well-defined.

Mathematically, all of this is summarised by saying that a process theory is precisely a *symmetric monoidal category* $(\mathbf{C}, \otimes, I)$ with the processes as its *morphisms*. Our diagrammatic rules correspond to the precise *graphical calculus* for reasoning in such categories (Selinger, 2011).

We will often wish to refer to some special kinds of processes. Processes with 'no input' in diagrams (and so formally with input object I) are called *states*, and can be thought of as 'preparations' of the physical system given by their output object:

Processes with no output, called *effects*, may be thought of as 'observations' we may record on our system. Finally, processes with neither input nor output are called *scalars*. It is common for theories to come with a *probabilistic* interpretation meaning that each of their scalars p correspond to a probability, or more generally an 'unnormalised probability' $p \in \mathbb{R}^+$, with $r \otimes s = r \cdot s$ for scalars and the empty diagram given by 1. In particular, the composition of a state with an effect



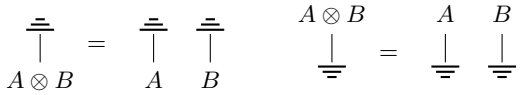
$\in \mathbb{R}^+$

corresponds to the 'probability' of observing the effect e in the state ρ. Such 'generalised probabilistic theories' are a major focus of study in the foundations of physics (J. Barrett, 2007).

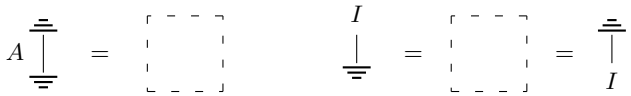
The theories we consider here will often come with further structure giving them a physical interpretation. Firstly, every object will come with a distinguished *discarding* effect depicted

which we think of as the process of simply 'throwing away' or 'ignoring' a physical system. Similarly, every object should come with a distinguished *completely mixed state* depicted as

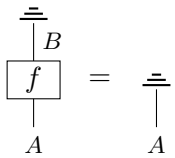
which corresponds to preparing the object in a maximally 'noisy' or 'random' state. These processes should satisfy



as well as



for all objects A, B. We then define a process f to be *causal* when it satisfies



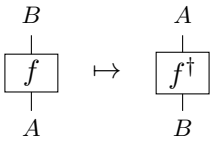
or similarly as *co-causal* if it preserves $\frac{\perp}{\overline{\overline{\overline{}}}}$. Discarding processes are in fact closely related to physical notions of causality; see for example (Coecke, 2014; Chiribella, D'Ariano, & Perinotti, 2010).

In such a probabilistic theory there is a unique process between any two objects, the *zero process* 0, such that composing any process via $\circ, \otimes$ with 0 always yields 0.

At times we will assume our process theory also comes with a way of describing how similar any two causal states are. This amounts to a choice of *distance function* on the set $\mathsf{St}_c(A)$ of causal states of each object A, providing a value $d(a, b) \in \mathbb{R}^+$ for each $a, b \in \mathsf{St}_c(A)$. Often this map d will satisfy the axioms of a metric, but this is not required.

Our main examples of process theories will come with a notable extra feature, though this will not be necessary for our approach. In many theories it is possible to 'reverse' any process, in that for any process f there is another $f^\dagger$ in the opposite direction. We

say a process theory has a *dagger* when it comes with such a mapping



which preserves composition and identity maps in an appropriate sense, and satisfies $f^{\dagger\dagger} = f$ for all f. The presence of a dagger is a common starting point in categorical approaches to quantum theory; see e.g. (Abramsky & Coecke, 2004; Selinger, 2007).

Let us now meet our main examples of process theories with the above features.

Example 3.2.1. *(**Classical probabilistic processes**) In the process theory $\mathbf{Class}$ of finite-dimensional probabilistic classical physics, the objects are finite sets $A, B, C, \ldots$ and the processes f from A to B are functions sending each element $a \in A$ to a 'unnormalised probability distribution' over the elements of B, i.e functions $f\colon A \times B \to \mathbb{R}^+$. Composition of f from A to B and g from B to C is defined by*

$$(g \circ f)(a, c) = \sum_{b \in B} f(a, b) \cdot g(b, c)$$

In this theory the trivial object is the singleton set $I = \{\star\}$, with $\otimes$ given by the Cartesian product $A \times B$ and $(f \times g)(a, c)(b, d) = f(a, b) \cdot g(c, d)$. This theory is probabilistic, with scalars $r \in \mathbb{R}^+$.

Here $\fatslash_A$ is the unique effect with $\fatslash_A(a) = 1$ for all $a \in A$. A process f is causal whenever it is stochastic, i.e. sends each element $a \in A$ to a (normalised) probability distribution over the elements of B. Applying the process $\fatslash$ to some output wire of a process corresponds to marginalising *over the set which is discarded.*

States of an object are '$\mathbb{R}^+$-distributions' over their elements, while causal states are normalised ones, i.e. probability distributions. The completely mixed state $\fatbslash_A$ is the uniform probability distribution, with $\fatbslash_A(a) = \frac{1}{|A|}$ for all $a \in A$. This theory also has a dagger by $f^\dagger(b, a) = f(a, b)$.

Similarly we define another process theory $\mathbf{Class}_{\mathrm{m}}$, in the same way, but with objects now being finite metric spaces *(A, d). Each object A now comes with a metric d on its underlying set, with $A \otimes B = A \times B$ having the product metric. For each object A we extend d to a metric d_W on probability distributions over A, i.e. causal states of A, called the* Wasserstein metric *or* Earth Mover's Distance *(EMD), definable e.g. by*

$$d_W(s, t) := \sup_f \{\sum_{a \in A} f(a) \cdot s(a) - \sum_{a \in A} f(a) \cdot t(a)\}$$

where the suprema is taken over all functions f satisfying $|f(a) - f(b)| \leq d(a, b)$ for all a, b. $\mathbf{Class}$ itself may be given a metric on causal states in the same way by taking each object A to have metric $d(a, b) = 1 - \delta_{a,b}$.

Example 3.2.2. *(**Quantum Processes**) In the process theory* $\mathbf{Quant}$ *the objects are finite-dimensional complex Hilbert spaces* $\mathcal{H}, \mathcal{K}, \ldots$ *and the processes from* $\mathcal{H}$ *to* $\mathcal{K}$ *are* completely positive maps $f\colon B(\mathcal{H}) \to B(\mathcal{K})$ *between their spaces of operators. Here* $I = \mathbb{C}$ *and* $\otimes$ *is the usual tensor product of Hilbert spaces and maps. States* ρ *of an object* $\mathcal{H}$ *may be identified with (unnormalised)* density matrices*, i.e. quantum states in the usual sense, as may effects. The effect* $\top$ *sends each operator* $a \in B(\mathcal{H})$ *to its* trace $\mathrm{Tr}(a)$*, and* $\bot$ *is the maximally mixed state on* $\mathcal{H}$*, with density matrix* $\frac{1}{\dim(\mathcal{H})} 1_{\mathcal{H}}$*. Here a process is causal precisely when it is trace-preserving, and the dagger is given by the Hermitian adjoint.*

Example 3.2.3. *(**Quantum-Classical Processes**) To combine* $\mathbf{Class}$ *and* $\mathbf{Quant}$ *we may use the theory* $\mathbf{CStar}$ *whose objects are finite-dimensional* C*-algebras $A, B, \ldots$ *and processes are completely positive maps* $f\colon A \to B$*, with* $\otimes$ *given by the standard tensor product,* $I = \mathbb{C}$ *and the dagger again by the Hermitian adjoint. Here* $\top$ *sends each element* $a \in A$ *to its trace* $\mathrm{Tr}(a) \in \mathbb{C}$*, while* $\bot$ *corresponds to the rescaling* $\frac{1}{d} 1$ *of the element* $1 \in A$*, where* $\mathrm{Tr}(1) = d$*. Each C*-algebra comes with a metric induced by its norm, providing a metric on states in the theory.*

$\mathbf{Class}$ *may be identified with the sub-theory of* $\mathbf{CStar}$ *containing the commutative algebras, and* $\mathbf{Quant}$ *with those of the form* $B(\mathcal{H})$ *for some Hilbert space* $\mathcal{H}$*. More general algebras are 'quantum-classical', being given by direct sums of quantum algebras.*

3.3. Decompositions

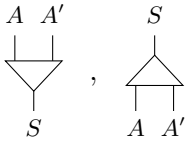
A central aspect of IIT is evaluating the level of integration of a process, and particularly of a state of some object. To do so we must compare the object in question against ways it may be *decomposed*, as follows.

Firstly, recall that a process f from A to B is an *isomorphism* when there is some (unique) f^{-1} from B to A for which $f^{-1} \circ f$ and $f \circ f^{-1}$ are both identities. We write $A \simeq B$ when such an isomorphism exists.

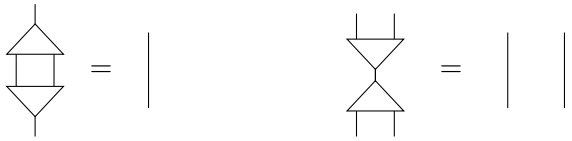
Definition 3.3.1. *In any process theory, a* decomposition *of an object* S *is a pair of objects* A, A' *along with an isomorphism* $S \simeq A \otimes A'$*.*

In a process theory with $\top, \bot$ *we will always consider decompositions whose isomorphisms are causal and co-causal. We also assume that decomposition isomorphisms preserve any distances between causal states.*

For short we often denote such a decomposition simply by (A, A') and depict its isomorphism and inverse by

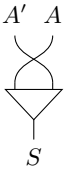


respectively. The fact that they form an isomorphism means that



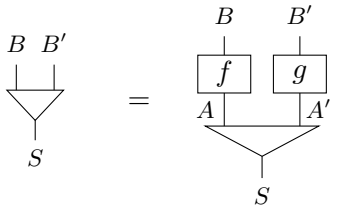
One can go on to develop a general study of decompositions in process theories. Here we just note some of the basics, for more see Appendix 3.A.

Firstly, any decomposition has an induced *complement* decomposition $(A, A')^\perp := (A', A)$, with isomorphism given by swapping its components:



All decompositions then satisfy $(A, A')^{\perp\perp} = (A, A')$. Moreover, any object always S always comes with *trivial decompositions* denoted $1 := (S, I)$ and $0 := (I, S)$ with $0 = 1^\perp$. Drawing either of their isomorphisms would just mean drawing a blank wire labelled by S.

It is also useful to note when two decompositions of an object are 'essentially the same'. We write $(A, A') \sim (B, B')$ and call both decompositions *equivalent* when there exists isomorphisms f, g with



(3.1)

In a theory with $\dividecontinuous, \dividedbarbelow$ we require moreover that f, g are causal and co-causal.

We write $\mathbb{D}(S)$ for the set of all equivalence classes of decompositions of S under $\sim$ (we will ignore the fact that in full generality each equivalence class may be a proper class rather than a set). Often we abuse notation and denote the members of simply by (A, A') instead of as equivalence classes $[(A, A')]_\sim$. It is easy to see that if two decompositions are equivalent then so are their complements, so that $(-)^\perp$ is well-defined on $\mathbb{D}(S)$.

Definition 3.3.2. *By a* decomposition set *of an object S in a process theory we mean a subset $\mathbb{D}$ of $\mathbb{D}(S)$ containing 1 and closed under $(-)^\perp$.*

Given any decomposition set $\mathbb{D}$ of S and any $(A, A') \in \mathbb{D}$, we define the *restriction* of

$\mathbb{D}$ to A via this decomposition to be the decomposition set

$$\mathbb{D}|_A := \left\{ \; \begin{array}{c} B \;\; C \\ \diagdown\!\diagup \\ A \end{array} \;\; | \;\; \exists \;\; \begin{array}{c} C \;\; A' \\ \diagdown\!\diagup \\ B' \end{array} \;\; \text{s.t.} \;\; \begin{array}{c} B \quad B' \\ \vert \quad \vert \\ \vert \; C \\ \diagdown\!\diagup \\ A \quad\quad A' \\ \diagdown\!\diagup\!\diagdown\!\diagup \\ S \end{array} \;\; \in \mathbb{D} \; \right\} \subseteq \mathbb{D}(A)$$

Intuitively $\mathbb{D}|_A$ consists of all decompositions of A which themselves can be extended to give a decomposition of S belonging to $\mathbb{D}$, via (A, A').

The most important examples of decomposition sets are the following.

Example 3.3.3. *Let S be an object with a given isomorphism*

$$S \simeq S_1 \otimes \cdots \otimes S_n$$

representing S as finite tensor of objects S_i which we may call elements*. This induces a decomposition set $\mathbb{D}$ of S whose elements correspond to subsets J of the elements. For any such subset, defining $S_J := \bigotimes_J S_j$ we have a decomposition $S \simeq S_J \otimes S_{J'}$ where J' is the set of remaining elements. Then $\mathbb{D}|_{S_J}$ contains a decomposition for each $K \subseteq J$ in the same way.*

Decompositions via elements as above are the only kinds appearing in classical or quantum IIT. However, more general ones allow us to treat systems which are not decomposable into any finite set of 'elementary' subsystems.

3.4. Systems

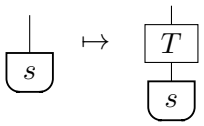
We now begin by seeing how each of the main components of IIT, or any 'generalised IIT' in the sense of (Kleiner & Tull, 2021), may be treated starting from any given process theory $\mathbf{C}$. The focus will be on a class of *systems*, as follows.

Definition 3.4.1. *By a* system type *we mean a triple $\underline{S} = (S, \mathbb{D}, T)$ consisting of an object S with a decomposition set $\mathbb{D}$ and a causal process*

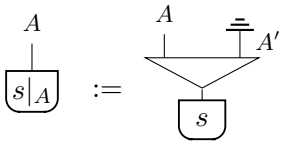
$$\begin{array}{c} S \\ \vert \\ \boxed{T} \\ \vert \\ S \end{array}$$

which we call its time evolution*. A* state *of $\underline{S}$ is simply a state of S in $\mathbf{C}$. We typically refer to a system type simply as a* system*.*

The set $\mathbb{D}$ specifies the ways in which we will decompose our underlying system when assessing integration. The process T is intended to describe the way in which states of the system evolve over each single 'time-step', via



In what follows it will be useful to be able to restrict any state s of our system to the components of any decomposition $(A, A') \in \mathbb{D}$ by setting

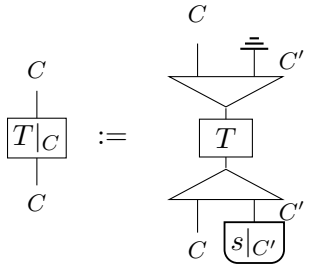


and defining $s|_{A'}$ similarly. We define the *trivial system* $\underline{I}$ to have object I, a single decomposition $1 = (I, I) = 0$, and time evolution being the identity.

3.4.1. Subsystems

There are several operations on systems one carries out in the context of IITs. The first is the taking of *subsystems*.

Definition 3.4.2. *For each object C belonging to some decomposition $(C, C') \in \mathbb{D}$, and each state s of $\underline{S}$, the corresponding* subsystem *of $\underline{S}$ is defined to be the system type $\underline{C}^s := (C, \mathbb{D}|_C, T|_C)$ with time evolution*



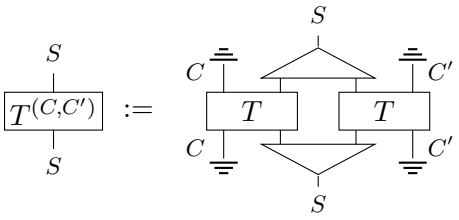
The above definition of $T|_C$ is from (Oizumi et al., 2014) and aims to capture the evolution of a state of C conditioned on the state of C' being $s|_{C'}$.

3.4.2. Cutting

A second important operation involves removing (some or all) causal connections between the two different components of a decomposition of a system. For any system $\underline{S} = (S, \mathbb{D}, T)$ and decomposition $(C, C') \in \mathbb{D}$, we should be able to form a new such *cut* system of the form

$$\underline{S}^{(C,C')} = (S, \mathbb{D}, T^{(C,C')})$$

with the new evolution $T^{(C,C')}$ removing some influence between these regions. The most straightforward form of cutting is a *symmetric cut*, in which both components are fully disconnected from each other, with evolution

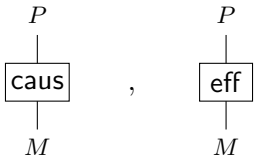


(3.2)

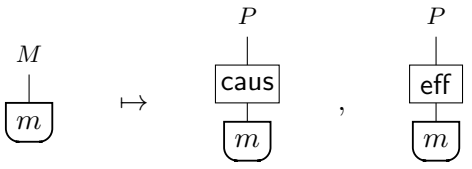
(where the triangle denotes $(C, C')^{\perp}$). However, later we will see that some IITs use additional structure to carry out alternative notions of system cut.

3.5. Cause and Effect

Central to any IIT is a notion of causal influence between any two possible subsystems of a system. These influences are captured in a pair of assignments called the *cause repertoire* and *effect repertoire* of the system. In IIT 3.0 these contain probability distributions describing how the present state of each subsystem constrains the past and future states of each other subsystem (Oizumi et al., 2014). For our purposes it suffices to note that such cause and effect repertoires amount to specifying a pair of processes

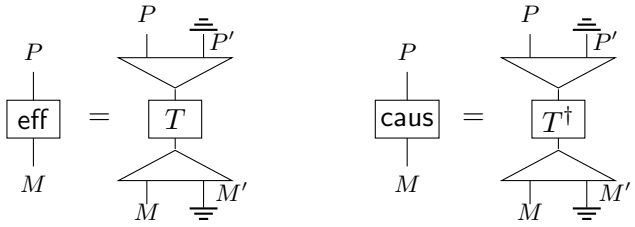


for each pair of underlying objects M, P of subsystems $\underline{M}, \underline{P}$ of $\underline{S}$ via some state s. In this setting M is typically called the 'mechanism' and P the 'purview', and the above processes should capture the way in which the current state m of M constrains the previous or next state of P, respectively. These constraints are captured by the pair of states of P given by plugging in the 'current' state m of M:



We will additionally require the processes caus, eff to be *weakly causal* in the sense that whenever the state m is causal then each of the above states must either be causal or 0.

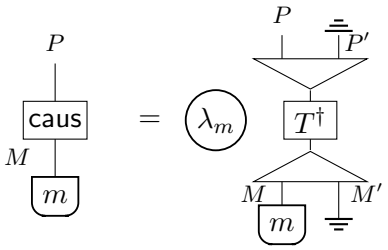
Example 3.5.1. *For any process theory (resp. with a dagger) there is a simple choice of effect (resp. cause) repertoire given by*



(3.3)

Note however that this definition of caus *may not be weakly causal in our above sense if* $T^\dagger$ *is not causal.*

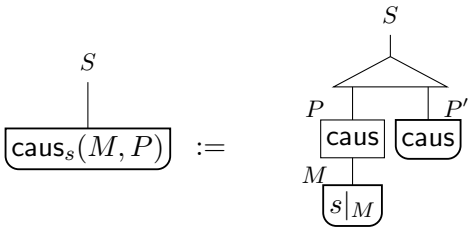
In a probabilistic process theory we should instead have that



(3.4)

where λ_m *is the unique* normalisation *scalar for the right-hand state, making it causal if it is non-zero (and being zero otherwise). It is not in general possible to define a process* caus *in terms of its action on states* m *in this way, but this is possible for example in* $\mathbf{Class}$, $\mathbf{Quant}$ *or* $\mathbf{CStar}$.

However the repertoires are specified, we will need to compare their values in a fixed state while varying P. To do so, for each state s of $\underline{S}$ and each such M, P we define the *cause repertoire at* s to be the state of S given by



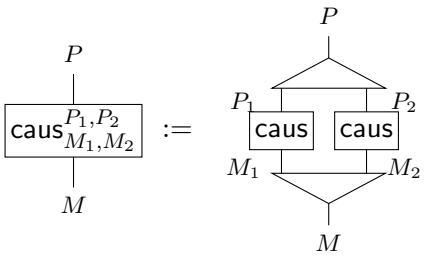
(3.5)

The features of this diagram have special names in (Oizumi et al., 2014); the right-hand caus state above, given by taking mechanism $M = I$, is called the *unconstrained* cause repertoire, and the whole process above $s|_M$ in the diagram is called the *extended cause repertoire* at M, P. Defining them in this way allows us to compare the repertoire values for varying M, P.

Similarly, $\mathsf{eff}_s(M, P)$, the *effect repertoire at* s, and the *unconstrained* and *extended effect repertoire* are all defined in terms of eff in the same way.

3.5.1. Decomposing repertoires

In an IIT we must assess how integrated each of these repertoire values are at a given state . This involves comparing the repertoires with how they behave under decomposing each of M and P. For any decompositions $(M_1, M_2) \in \mathbb{D}|_M$ of M and $(P_1, P_2) \in \mathbb{D}|_P$ of P, the *decomposed* cause repertoire process is defined by



(3.6)

We then define the state $\mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M,P)$ just like (3.5) but replacing caus with the process (3.6). We decompose the effect repertoire in just the same way in terms of eff.

3.6. Generalised IITs

In summary, let $\mathbf{C}$ be a process theory coming with the features $\dot{+}, \dot{-}, d$ of Section 3.2. To define an integrated information theory we must specify:

1. a class $\mathbf{Sys}$ of system types, closed under subsystems;
2. a definition of system cuts, under which $\mathbf{Sys}$ is closed;
3. a choice of weakly causal processes caus, eff between the underlying objects M, P of each pair of subsystems $\underline{M}, \underline{P}$ via some state s, of any system $\underline{S}$.

More precisely, this provides the *data* of a generalised integrated information theory in the sense of (Kleiner & Tull, 2021). From this data we may now use the *IIT algorithm* from (Oizumi et al., 2014) to calculate the usual objects of interest in IIT.

3.6.1. The IIT Algorithm

We now briefly summarise this algorithm as treated in the general setting in (Kleiner & Tull, 2021), to which we refer for more details. Let us fix a 'current' state s of a system $\underline{S}$. Firstly, the level of *integration* of each value of the cause repertoire is defined by

$$\phi(\mathsf{caus}_s(M,P)) \;:=\; \min d(\mathsf{caus}_s(M,P)\;,\; \mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M,P)) \tag{3.7}$$

where the minima is taken over all pairs of decompositions of M, P which are not both trivial, i.e. equal to 1. [2] The integration level $\phi(\mathsf{eff}_s(M,P))$ is defined similarly in terms of eff.

[2] When $\mathsf{caus}_s(M,P) = 0$ we alternatively set $\phi = 0$.

For each choice of mechanism M, its *core cause* P^c and *core effect* P^e are the purviews P with maximal ϕ values for $\mathsf{caus}, \mathsf{eff}$ respectively. The minima of their corresponding ϕ values is then denoted by $\phi(M)$. We then associate to M and object called its *concept* $\mathbb{C}(M)$, essentially defined as the triple

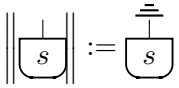
$$(\mathsf{caus}_s(M, P^c), \mathsf{eff}_s(M, P^e), \phi(M))$$

More precisely, in (Kleiner & Tull, 2021), $\mathbb{C}(M)$ is given by the pair of above repertoire values with each 'rescaled' by $\phi(M)$.

The tuple $\mathbb{Q}(s)$ of all these concepts, for varying M, is called the *Q-shape* $\mathbb{Q}(s)$ of the state s. The collection of all possible such tuples is denoted $\mathbb{E}(\underline{S})$. The level of integration of $\mathbb{Q}(s)$ is calculated similarly to (3.7) by considering all possible cuts of the system. The subsystem $\underline{M}$ of $\underline{S}$ whose Q-shape is itself found to be most integrated is called the *major complex*. Rescaling this Q-shape $\mathbb{Q}(\underline{M}, s|_M)$ according to its level of integration, and using an embedding $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$ we finally obtain a new element $\mathbb{E}(s) \in \mathbb{E}(\underline{S})$.

The claim of an IIT with regards to consciousness is that $\mathbb{E}(\underline{S})$ is the space of all possible conscious experiences of the system $\underline{S}$, and that $\mathbb{E}(s)$ is the particular experience attained when it is in the state s, with intensity $\Phi(s) := \| \mathbb{E}(s) \|$.

Remark 3.6.1. *Let us make explicit how the specification of 1, 2, 3 above provides the data of an IIT in the sense of (Kleiner & Tull, 2021). The system class of the theory is* $\mathbf{Sys}$*, and* $\mathsf{caus}_s(M, P), \mathsf{eff}_s(M, P)$ *and their decompositions are as outlined in Section 3.5.1. When* $\mathbf{C}$ *is probabilistic and has distances* $d(a, b)$ *defined for* arbitrary *states* a, b *of an object* A*, we may define the space of* proto-experiences $\mathbb{PE}(\underline{S})$ *of a system* $\underline{S}$ *to be simply its set of states, with*



However, if d *is only defined on causal states, as in classical IIT, to follow the algorithm from (Kleiner & Tull, 2021) one must instead set* $\mathbb{PE}(\underline{S}) := \mathsf{St}_\mathsf{c}(S) \times \mathbb{R}^+$ *as in Example 3 of (Kleiner & Tull, 2021). For either choice, for any subsystem* $\underline{M}$ *of* $\underline{S}$ *we obtain an embedding* $\mathbb{PE}(\underline{M}) \hookrightarrow \mathbb{PE}(\underline{S})$ *by composing alongside* $\frac{}{=} M^\perp$*, and this can be seen to provide a further embedding* $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$*.*

3.7. Examples

Let us now meet several examples of IITs defined from process theories.

3.7.1. Generic IITs

Let $\mathbf{C}$ be any process theory coming with the features outlined in Section 3.2, including a dagger on processes. We define a generalised IIT denoted $\mathsf{IIT}(\mathbf{C})$ by taking as systems

all tuples $\underline{S} = (S, \mathbb{D}, T)$ of an object S in $\mathbf{C}$ along with a causal process T and a decomposition set $\mathbb{D}$ induced by a single isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ in terms of elements S_i, as in Example 3.3.3. As before each partition of these elements gives a decomposition of S. We define system cuts to be symmetric as in (3.2) and the repertoires in the straightforward sense of (3.3).

Remark 3.7.1. *We can extend this example in to ways. Firstly we may allow systems $\underline{S}$ to come with arbitrary finite decomposition sets $\mathbb{D}$ of S. Secondly, we may extend the definition to theories without daggers by instead simply requiring each system $\underline{S}$ to come with a process T^- describing 'reversed time evolution', and then define the cause repertoire by replacing $T^\dagger$ with T^-.*

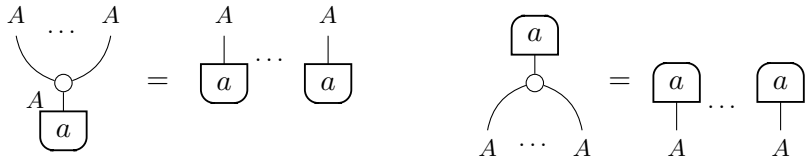
3.7.2. Classical IIT

The 'classical' IIT version 3.0 of Tononi and collaborators (Oizumi et al., 2014) is built on the process theory $\mathbf{Class}_\mathsf{m}$. As such a toy model of the theory is provided by $\mathsf{IIT}(\mathbf{Class}_\mathsf{m})$. However IIT 3.0 itself differs from this theory, using some more specific features of the process theories $\mathbf{Class}$ and $\mathbf{Class}_\mathsf{m}$ which we now describe.

Firstly, note that in these classical process theories, for each object A, each element $a \in A$ corresponds to a unique state given by the point distribution at a, as well as a unique effect, namely the map sending a to 1 and all other elements of A to 0. We denote this state and effect both simply by a.[3]

Any process f from A to B is determined entirely by its compositions with these special states and effects since plugging in such a state a and effect b yields its value $f(a, b)$.

Another special feature of these classical process theories is that each object A comes with a distinguished *copying* process from A to $A \otimes \cdots \otimes A$, for any number of copies of A, as well as a *comparison* process in the opposite direction. We denote and define these respectively by the rules

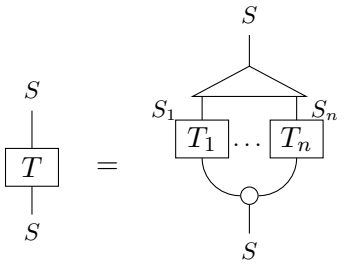


for all $a \in A$. Abstractly, these operations form a canonical commutative *Frobenius algebra* on each object, and there is no such canonical algebra on each object in $\mathbf{Quant}$ due to the *no-cloning* theorem (Coecke, Pavlovic, & Vicary, 2013). We may now describe IIT 3.0 itself as follows.

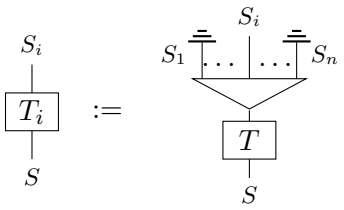
[3] Typically these are the only kinds of 'state' considered, e.g. in (Oizumi et al., 2014) and even in our related article (Kleiner & Tull, 2021). In contrast here the term 'state' would include all distributions over A, i.e. all states of the process theory $\mathbf{Class}_\mathsf{m}$.

3.7.2.1. Systems

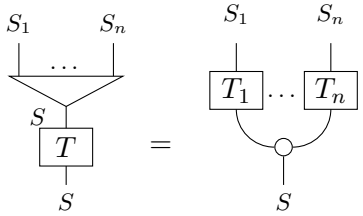
In this theory systems are defined similarly to IIT($\mathbf{Class}_{\mathrm{m}}$), being given by a finite metric space S given as a product of elements $S \simeq \bigotimes_{i=1}^{n} S_i$, along with a causal (i.e. stochastic) evolution T on S. Additionally in (Oizumi et al., 2014) each evolution T is required to satisfy *conditional independence*, which states that for all $s, t \in S$, with $t = (t_1, \ldots, t_n)$ for some $t_i \in S_i$ we have



where for each element S_i we define the process T_i by



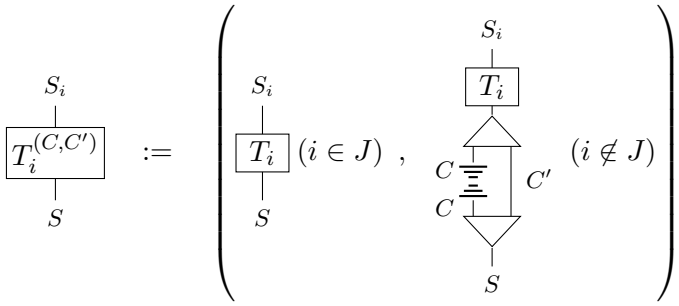
having depicted the isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ by the triangle above. In other words, conditional independence states that the probabilities for the next state of each element S_i are independent. Equivalently, T must satisfy



3.7.2.2. Cuts

Rather than our earlier symmetric cuts, the system cuts used in IIT 3.0 are *directional*. For any decomposition (C, C') of S with $C = \bigotimes_{j \in J} S_j$ for some subset of notes indexed by $J \subseteq \{1, \ldots, n\}$, we define the cut evolution $T^{(C,C')}$ using conditional independence

by setting

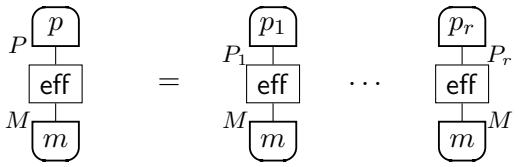


In other words, in the cut system all causal connections $C \to C'$ are replaced by noise, while all those into C remain intact.

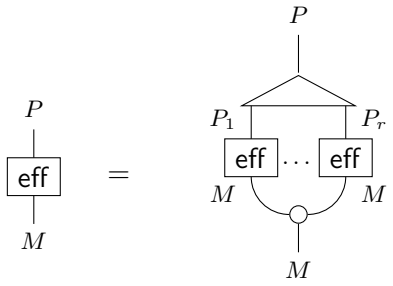
3.7.2.3. Repertoires

Let us now define the processes caus, eff between a pair of objects M and P, with $M = \bigotimes_{i=1}^{k} M_i$ and $P = \bigotimes_{j=1}^{r} P_j$ for some subsets $\{M_1, \ldots, M_k\}$ and $\{P_1, \ldots, P_r\}$ of elements of the system.

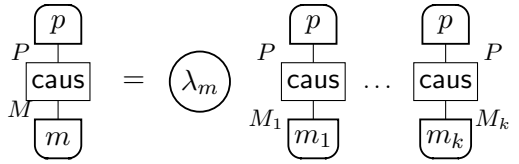
We begin with eff. When P is simply a single element P_j, eff is defined exactly as in (3.3). For more general P we define eff to again satisfy a form of conditional independence, so that



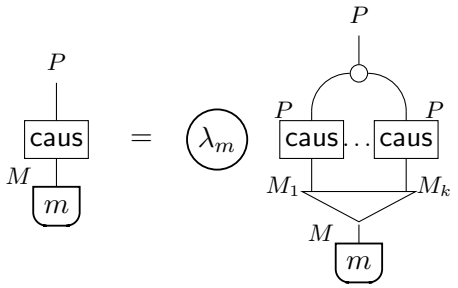
for all $m \in M, p = (p_1, \ldots, p_r) \in P$. Equivalently, we have that



In a similar fashion, whenever M is a single element M_i we define caus from M to P as in (3.4), while for more general M we require that



for all $m = (m_1, \ldots, m_k) \in M$ and $p \in P$, where λ_m is the normalisation scalar making $\mathsf{caus} \circ m$ a causal state (probability distribution) if it is non-zero, or $\lambda_m = 0$ otherwise. Equivalently, this means that



for each $m \in M$. This concludes the data of classical IIT.

3.7.3. Quantum IIT

Zanardi, Tomka and Venuti have proposed a quantum extension of classical IIT (Zanardi et al., 2018). In fact it is comparatively much simpler to describe in our approach, being precisely the theory $\mathrm{IIT}(\mathbf{Quant})$.

Explicitly, systems in this theory are given by finite-dimensional complex Hilbert spaces $\mathcal{H}$ along with a given decomposition into elements $\mathcal{H} \simeq \bigotimes_{i=1}^{n} \mathcal{H}_i$ and a completely positive trace-preserving map T on $B(\mathcal{H})$. States and repertoire values are given by density matrices ρ. In this theory each Q-shape $\mathbb{Q}(\rho)$ may be encoded as a single positive semi-definite operator on the space $(\mathbb{C}^2)^{\otimes n} \otimes \mathbb{C}^2 \otimes \mathcal{H}$, as discussed in (Zanardi et al., 2018).

3.7.4. Quantum-Classical IIT

We may now define a version of *quantum-classical IIT* as $\mathrm{IIT}(\mathbf{CStar})$. This synthesizes quantum IIT with the toy version $\mathrm{IIT}(\mathbf{Class}_{\mathrm{m}})$ of classical IIT, containing both kinds of systems. In future it would be desirable to synthesise quantum IIT with IIT 3.0 proper. Since the latter relies on the presence of copying maps, this may be achievable using the more general notion of a *leak* on a C*-algebra (Selby & Coecke, 2017).

3.8. Outlook

In this article we have taken first steps to show how Integrated Information Theory, and its generalisations to other domains of physics, may be studied categorically. There are many avenues for future work.

Firstly, we have so far made no requirements on the cause and effect repertoire processes $\mathsf{caus}, \mathsf{eff}$. To be fit for their name these processes should be required to satisfy axioms which ensure they have a causal interpretation, ideally determining them uniquely within any given process theory. Monoidal categories provide a natural setting for the

study of causality, a major contemporary topic in the foundations of physics (Kissinger & Uijlen, 2017).

At a higher level, it seems natural for the class of systems $\mathbf{Sys}$ of a generalised IIT to itself form a category. The theory itself should then give a functor into another category $\mathbf{Exp}$ of (spaces of) phenomenal experiences; a formalization of the latter is for example given in (Kleiner & Tull, 2021).

Making IIT functorial in this way will likely involve modifying it to be more natural from a categorical perspective. Developing a useful notion of integration applicable to any monoidal category may also help to resolve mathematical problems of the IIT algorithm, for example its relying on the unique existence of core purviews which are not guaranteed

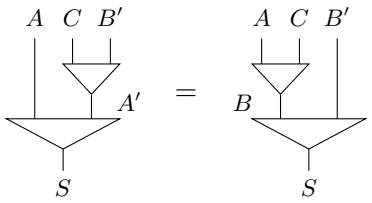
Appendix

3.A. Decompositions and Integration

Here we briefly mention a few further results about decompositions of objects in process theories; we leave a detailed study of their properties to future work.

Our earlier definition of $\mathbb{D}|_A$ was based on an idea of one decomposition as being 'contained in' another. Let us make this precise.

Definition 3.A.1. *Let S be an object in a process theory and (A, A'), (B, B') two decompositions. We write that $(A, A') \preceq (B, B')$ whenever there exists an object C and decompositions (A, C) of B and (B', C) of A' such that*

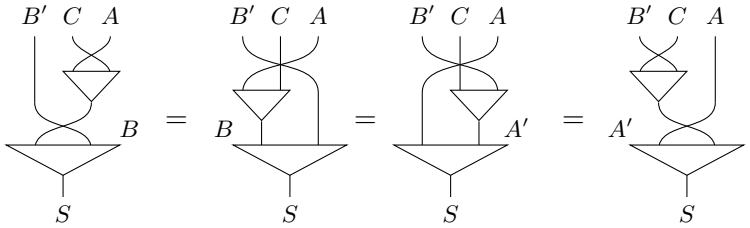


(3.8)

Intuitively, this states that A is contained in B (as is B' within A') in a way compatible with these decompositions.

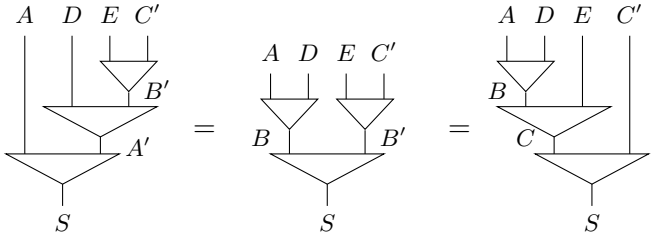
Lemma 3.A.2. *Let S be an object in a process theory. Then $\preceq$ forms a pre-order on the set of decompositions of S, with top element 1 and bottom element 0, and $(-)^\perp$as an involution.*

Proof. We always have $(A, A') \preceq (A, A')$ by taking $C = I$ and using the decompositions 1 and 0 on A in (3.8). Similarly $(A, A') \preceq 1$ by taking $C = A'$. To see that $(-)^\perp$ is an involution, suppose that $(A, A') \preceq (B, B')$ as above. Then we have $(B, B')^\perp \preceq (A, A')^\perp$ since



Hence we always have $0 = 1^\perp \preceq (A, A')$ for all (A, A'). For transitivity, note that

whenever $(A, A') \preceq (B, B') \preceq (C, C')$ via some respective objects D, E then we have

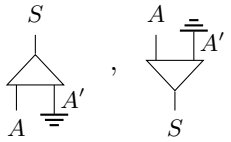


so that $(A, A') \preceq (C, C')$ via the above decompositions $(D \otimes E, C')$ of A' and $(A, D \otimes E)$ of C. ■

Recall that in any category, a *sub-object* of an object A is an (isomorphism class of a) monomorphism $m \colon M \to A$. It is *split* when $e \circ m = \mathrm{id}_M$ for some e. The sub-objects of A form a partial order $\mathrm{Sub}(A)$.

Lemma 3.A.3. *In any process theory with $\top, \bot$, for any object S:*

1. *Any decomposition (A, A') of S makes A a split sub-object of S via*

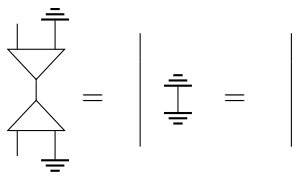


(3.9)

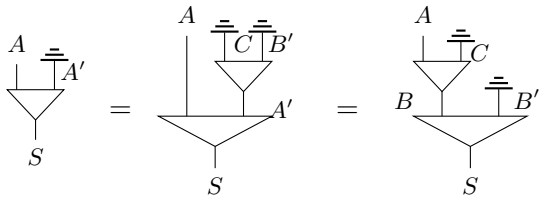
   Moreover if $(A, A') \preceq (B, B')$ then $A \leq B$ in $\mathrm{Sub}(S)$.

2. *$\preceq$ restricts to a partial order $\leq$ on $\mathbb{D}(S)$, again with top element 1, bottom 0 and involution $(-)^\perp$.*

Proof. 1: We have

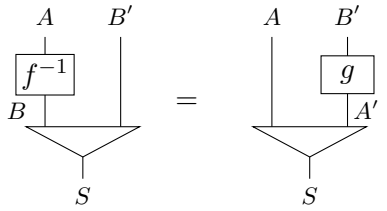


If $(A, A') \preceq (B, B')$ then the splitting for A factors over that for B since:



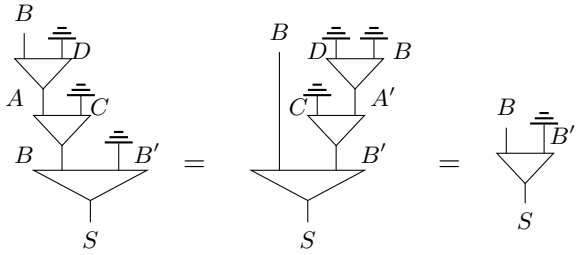
It follows that $A \leq B$ in $\mathrm{Sub}(S)$.

2: We need to show that any two decompositions (A, A') and (B, B') are equivalent under $\preceq$ precisely when they are equivalent in the sense of (3.1). Firstly, if there exists causal and co-causal isomorphisms f, g making (3.1) hold, then we have

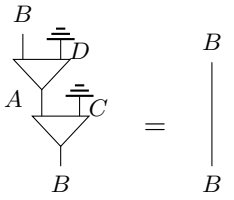


Viewing f^{-1} and g as decompositions (A, I) of B and (I, B') of A', respectively, this gives that $(B, B') \preceq (A, A')$. Then $(A, A') \preceq (B, B')$ holds similarly.

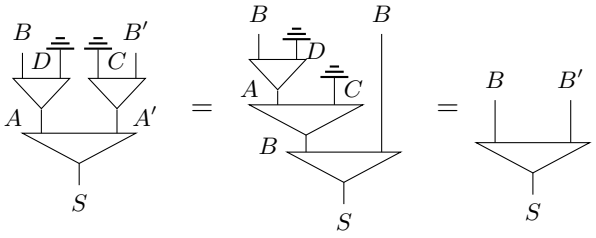
Conversely, if $(A, A') \preceq (B, B') \preceq (A, A')$, via respective objects C, D then



Since the right-hand map is an epimorphism by the first part, this gives that



Dually, composing in the other order gives the identity on A, making these causal and co-causal isomorphisms $A \simeq B$. Similarly we obtain such isomorphisms $A' \simeq B'$. Then we have



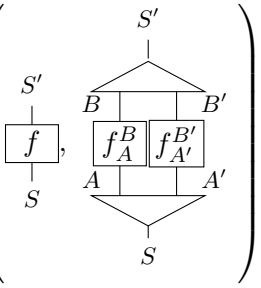
as required. Now 2 follows since any pre-order restricts to a partial order on its set of equivalence classes, and so $\preceq$ becomes a partial order $\leq$ on $\mathbb{D}(S)$. It is easy to see that the earlier properties of $1, 0, (-)^{\perp}$ carry over to $\leq$. ■

3.A.1. Integration

Let us briefly allude to how integration may generally be studied and quantified using decomposition sets.

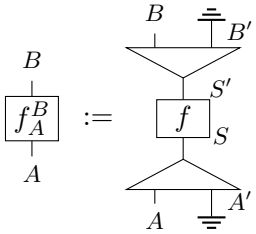
Suppose we have objects S, S' with given decomposition sets $\mathbb{D}, \mathbb{D}'$ and for each $(A, A') \in \mathbb{D}$ and $(B, B') \in \mathbb{D}'$ a process f_A^B from A to B. We denote $f_S^{S'}$ simply by f. Whenever we have a given distance function d on the set of processes from S to S', we may define the level of *integration* of the family $(f_A^B)_{A,B}$ as

$$\phi(f) := \min_{\mathbb{D}\times\mathbb{D}'} d \left(\quad , \quad \right)$$



where we exclude the top element $(1, 1)$ of $\mathbb{D} \times \mathbb{D}'$ in the minimisation.

Example 3.A.4. *Given any process f from S to S' we may define such a family $(f_A^B)_{A,B}$ with $f_S^{S'} = f$ by setting*



Example 3.A.5. *Our earlier description of the IIT algorithm precisely includes evaluating the integration level of each of the families of processes $(\mathsf{caus})_{M,P}$ and $(\mathsf{eff})_{M,P}$ using the state-dependent distance*

$$d_m \left(\boxed{f}^{P}_{M}, \boxed{g}^{P}_{M} \right) := d \left(\boxed{f}^{P}_{\boxed{m}}, \boxed{g}^{P}_{\boxed{m}} \right)$$

where $m = s|_M$ and d is the distance on $\mathsf{St}(S)$.

4. Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy

Sean Tull, Johannes Kleiner, Toby St Clere Smithe[1]

4.1. Introduction

Predictive processing (PP) is a framework for modelling cognition and adaptive behaviour in both biological and artificial systems (Wiese & Metzinger, 2017; Hohwy, 2020). A prominent sub-field is the programme of *Active Inference*, developed by Friston and collaborators (Smith, Friston, & Whyte, 2022; Parr et al., 2022; K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017; Sajid, Ball, Parr, & Friston, 2021), which aims to provide a unified understanding of cognition and action which can be applied at many levels, from a single neuron to an entire brain or organism. More specifically, active inference gives a proposal for how a cognitive agent represents its own beliefs about the world, how it updates these beliefs in light of new observations, and how it chooses the actions it takes, with the latter ultimately leading to new observations.

Central to the framework is that an agent possesses a *generative model* which explains its observations causally in terms of both hidden states of the world and its own actions. Note that this model is internal to the agent, and typically distinct from the 'true' causal process in the world which produces the observations. After receiving an observation, the agent may *update* this generative model to determine likely hidden states

[1]Accepted to the 6th International Conference on Applied Category Theory (ACT2023) as: Tull, S., Kleiner, J., & Smithe, T. S. C. (2023). Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy. (Tull et al., 2023)

which caused the observation (the process of perception) and choose its actions (the process of planning). In active inference, both forms of updating are carried out together through a form of approximate Bayesian inference, by minimising a quantity known as *free energy* (K. Friston, Kilner, & Harrison, 2006; K. Friston, 2010).

While active inference seeks a principled account of cognition, at present its formalisation can seem fairly complex, and there are various aspects which do not follow immediately from simply applying the definitions to a given generative model. Conceptually clear formal accounts of the framework would be desirable to simplify the theory and address these issues, as well as for applications within AI.

One hope for such a formal account would be for it to be both *compositional* and *graphical*. Indeed the generative models in PP are highly structured, often given as 'hierarchical models' (De Vries & Friston, 2017) which are best represented diagrammatically in terms of probabilistic graphical models such as Bayesian networks. While there has been support for, and steps towards, a graphical account of active inference (K. J. Friston, Parr, & de Vries, 2017), so far the graphical aspects only formally describe the structure of a generative model, while other aspects such as updating and free energy are still treated through traditional probabilistic calculations, and only informally in diagrams.

Recently however, fully formal diagrammatic methods have been developed for both describing Bayesian networks and carrying out probabilistic reasoning about them. These approaches are based on (monoidal) *category theory* and its associated graphical language of *string diagrams* (Piedeleu & Zanasi, 2023). Category theory has been applied across the sciences as a general mathematics of interacting *processes*, including within probability theory (Coecke & Spekkens, 2012; Cho & Jacobs, 2019), causality (Jacobs, Kissinger, & Zanasi, 2019; Fritz & Klingler, 2023; Lorenz & Tull, 2023), game theory (Ghani, Hedges, Winschel, & Zahn, 2018), machine learning (Fong, Spivak, & Tuyéras, 2019; Shiebler, Gavranović, & Wilson, 2021), quantum computing (Abramsky & Coecke, 2004) and natural language processing (S. Clark, Coecke, & Sadrzadeh, 2008). In particular a major ongoing development is in the study of probabilistic processes in terms of *cd-categories* (and 'Markov categories'), which allow one to carry out probabilistic reasoning entirely through string diagrams (Fritz, 2020).

In this work we give a full categorical account of predictive processing and active inference in terms of string diagrams, interpreted in cd-categories. In doing so we aim to give a conceptually clear account of the main features of the framework: generative models, Bayesian updating (including with soft observations), perception, action planning, and their combination in active inference, and both *variational* and *expected* free energy.

A highlight is a fully graphical derivation of the well-known formula for active inference in terms of minimisation of free energy. While this is a central result within active inference, its usual justification is more heuristic in nature. Here we instead derive the free energy formula purely graphically from a diagrammatic account of active inference itself, providing what we argue is the most transparent account of this result known so far.

The categorical perspective also naturally leads us to consider more novel aspects

of active inference. These include the definition of *open generative models* (essentially from (Lorenz & Tull, 2023)) which are generative models coming with 'inputs', allowing them to serve as the building blocks of an overall generative model.

We also introduce a notion of variational free energy for open models which allows us to establish the desirable property that free energy is *compositional*. Namely, a system with an overall generative model composed from sub-models may minimise global VFE by minimising VFE locally within each component. This is a crucial fact in order to apply free energy as proposed to all levels of a system, say from a whole brain down to its individual neurons.

Overall, we hope that our diagrammatic accounts of PP can provide a conceptually clear view of the framework, and also a natural language for reasoning within it. Indeed, as argued for example in (Lorenz & Tull, 2023) and elsewhere (Jacobs et al., 2019) diagrams in cd-categories provide a natural way to both represent causal (generative) models, as well as reason about them. As we demonstrate here they are also natural for describing the structure of active inference, including free energy. Aside from aiming to provide a helpful graphical language for those familiar with active inference, we conversely hope that this article may provide a succinct introduction to PP for those already familiar with string diagrams and categorical reasoning.

Further motivations Though primarily a framework for cognition, various proposals have been put forward for how predictive processing may be related to *consciousness* (Wiese & Metzinger, 2017). In previous work, two of the authors developed a categorical account of the *Integrated Information Theory* of consciousness, again essentially using cd-categories (Tull & Kleiner, 2021; Kleiner & Tull, 2021) and based on the work here we hope to give a categorical account of how consciousness may be accounted for within PP (Deane, 2021; Hohwy & Seth, 2020). We also see this work as a piece of the programme of *Compositional Intelligence*, which explores how categorically structured models and processes can be applied to (artificial) intelligence. Specifically, PP may be seen as a proposal for how compositional intelligence manifests in biology; that is, how biological systems may employ compositionality to carry out intelligent and adaptive behaviour.

Active inference can also be understood as an alternate proposal to reinforcement learning (RL) for how agents can learn adaptive behaviour, and shares similar features including the role of probabliistic models and inference (Tschantz, Millidge, Seth, & Buckley, 2020). It differs from conventional RL by replacing an explicit reward function with the aim of maximizing evidence for a probabilistic model, where the agent's preferences are now encoded in the model's prior distribution (K. J. Friston, Daunizeau, & Kiebel, 2009).

Related work This work can be seen as a part of the growing field of 'categorical cybernetics' (Smithe, 2021b; Capucci, Gavranović, Hedges, & Rischel, 2021), including previous work from one of the authors on compositional accounts of Bayesian updating (Smithe, 2020) and of active inference in terms of 'statistical games' (Smithe, 2021a,

2022). It differs from previous works by directly formalising the active inference framework itself, and by working explicitly graphically within the simple string-diagrammatic setting of cd-categories, with the aim of supplying a simple abstract characterization of active inference agents.

In this way the work is a part of a general movement in applying string diagrams in cd-categories to probability theory and causal reasoning. A categorical account of Bayesian inversion was first given by Coecke and Spekkens in (Coecke & Spekkens, 2012), and then within cd-categories by Cho and Jacobs (Cho & Jacobs, 2019), with further developments in categorical probability by Fritz (Fritz, 2020). Our diagrammatic account of generative models is precisely that given for causal models in part by one of the authors in (Lorenz & Tull, 2023), which builds on the earlier categorical treatments of (causal) Bayesian networks by Fong (Fong, 2013), Jacobs et al (Jacobs et al., 2019) and others e.g. (Fritz & Klingler, 2023). Indeed, as an agent's explanation for the observations it receives from the world, a generative model is ultimately a causal model (Pearl, 2009), though this is not often stressed in the literature.

The two forms of soft Bayesian updating treated here we first studied by Jacobs in (Jacobs, 2019). The specific treatment of conditioning in cd-categories used here is from (Lorenz & Tull, 2023). Cd-categories with (non-unique) conditionals have also been recently studied as 'partial Markov categories' in (Di Lavore & Román, 2023), along with both notions of updating. Our treatment of free energy refers to the KL divergence of distributions; we note that an axiomatic treatment of Markov categories coming with divergences on their morphisms has recently been given by Perrone in (Perrone, 2023).

Within active inference itself, graphical aspects have been increasingly prominent, with discussion of the 'graphical brain' in (K. J. Friston et al., 2017). In such works it is argued that one may describe models as (non-directed) Forney Factor Graphs (FFGs) (De Vries & Friston, 2017). However, generative models are inherently directed, going from states to observations with the other direction being intractable to compute exactly. Thus it is more natural to treat models using (generalisations of) directed Bayesian networks. Nonetheless we note though that FFGs derived from a model still have a role when minimising VFE via 'message passing' algorithms (Parr, Markovic, Kiebel, & Friston, 2019).

Interestingly, while a Bayesian network is typically depicted as a DAG (with only the variables labelled), one may argue the diagrams in active inference have been naturally 'converging' on their string diagrammatic representation, which also includes labels on the channels; see Figure 1. We claim that the advantage of string diagrams beyond DAGs is in allowing one to both represent and reason about the model in the same formalism.

We note that the diagrams in PP are at times only semi-formal, including aspects such as the free energy which are not strictly part of the generative model. One may see this work as a step towards the shared goal of representing formally all aspects of PP within one language of diagrams.

Structure of the article We begin in Section 4.2 by introducing cd-categories and their diagrammatic account of probability theory. We then apply these to introduce from

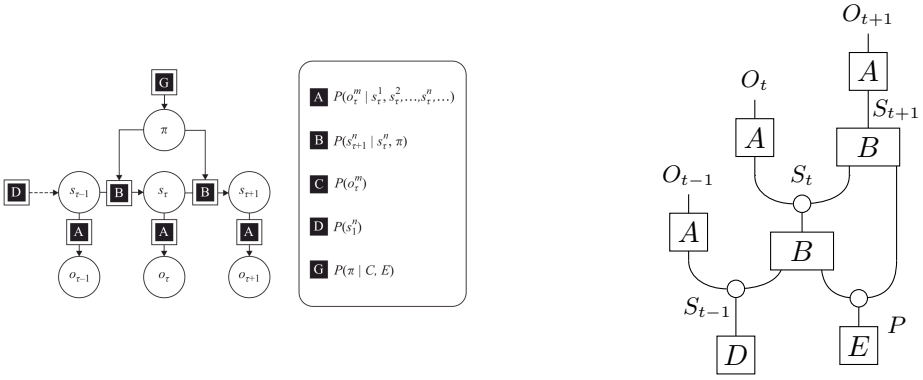


Figure 4.1.1.:
A generative model diagram from the recent book *Active Inference* by Parr, Pezzulo and Friston (Parr et al., 2022) (left) and the equivalent string diagram representation (right) (though replacing the informal EFE term G with the prior E).

scratch the key aspects of PP: generative models as Bayesian networks, and their generalisation to open generative models in a cd-category (Section 4.3), (Bayesian) updating of generative models from observations (Section 4.4), perception and planning (Section 4.5) and their combination in exact active inference (Section 4.6). We then discuss free energy (Section 4.7) and give a graphical derivation of active inference via free energy minimisation (Section 4.8). In Section 4.9 we then introduce free energy for open models using a graphical formalism of 'log-boxes' and use this to establish the compositionality property of free energy. Finally we discuss future work in Section 4.10.

Acknowledgements We thank Robin Lorenz for helpful discussions and development of the treatment of causal models used here. This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation. Sean Tull would also like to thank Quantinuum for their generous support in this research. Johannes Kleiner would like to thank the Mathematical Institute of the University of Oxford for hosting him while working on this research.

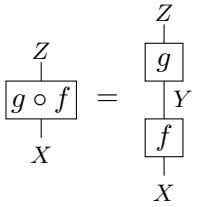
4.2. Categorical Setup

Let us begin by introducing the graphical treatment of probabilistic processes in terms of *string diagrams*, developed by numerous authors (Coecke & Spekkens, 2012; Cho & Jacobs, 2019; Fritz, 2020). Formally, these correspond to working in a 'monoidal cat-

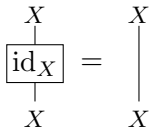
egory' or more specifically a 'cd-category', but in practice one may avoid mathematical details and simply work with the diagrams themselves. Though cd-categories are very general, in this article it suffices to consider the category $\mathbf{Mat}_{\mathbb{R}^+}$ of $\mathbb{R}^+$ valued finite matrices, introduced shortly in Example 4.2.1.

A *category* $\mathbf{C}$ consists of a collection of *objects* $X, Y, \ldots$ and *morphisms* or *processes* $f\colon X \to Y$ between them, which we can compose in sequence. In string diagrams we depict an object X as a wire and a morphism $f\colon X \to Y$ as a box with lower input wire X and upper output wire Y, read from bottom to top.

Given another morphism $g\colon Y \to Z$ we can compose them to yield a morphism $g \circ f\colon X \to Z$, depicted as:



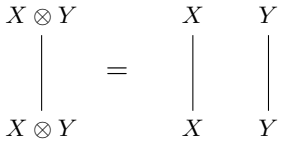
Each object X also comes with an *identity* morphism $\mathrm{id}_X\colon X \to X$ depicted as a blank wire:



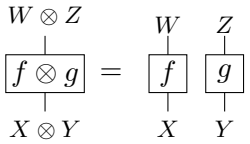
The identity leaves any morphism alone under composition, that is $\mathrm{id}_Y \circ f = f = f \circ \mathrm{id}_X$ for any $f\colon X \to Y$.

Formally, a *symmetric monoidal category* $(\mathbf{C}, \otimes, I)$ is a category $\mathbf{C}$ with a functor $\otimes\colon \mathbf{C} \times \mathbf{C} \to \mathbf{C}$, and natural transformations which express that $\otimes$ is suitably associative and symmetric, with a distinguished unit object I (Coecke, 2006). All of these aspects however are expressed most simply in diagrams.

Firstly, the *tensor* $\otimes$ allows us to compose any pair of objects X, Y into an object $X \otimes Y$, depicted by placing wires side-by-side.

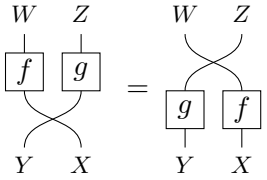


Given morphisms $f\colon X \to W$ and $g\colon Y \to Z$ we can similarly form their 'parallel composite' $f \otimes g\colon X \otimes Y \to W \otimes Z$ as below.

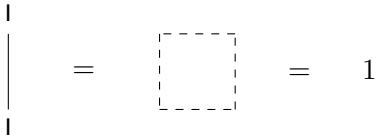


In text we will at times omit the tensor symbols and write e.g. 'f from X to Y, Z' or $f\colon X \to Y, Z$ in place of $f\colon X \to Y \otimes Z$.

The tensor is symmetric so we can 'swap' pairs of wires past each other, such that swapping twice returns the identity, and boxes carry along the swaps as below.

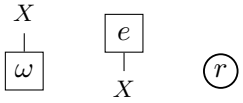


We also have a distinguished *unit object* I whose identity morphism we depict simply as empty space, and denote by 1.



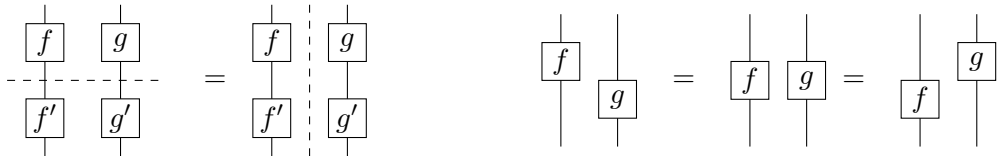
(4.1)

Intuitively, tensoring any object by the unit simply leaves it invariant. The unit allows us to consider morphisms with 'no inputs' and/or 'no outputs' in diagrams. A morphism $\omega\colon \mathsf{I} \to X$ is called a *state* of X, depicted with no input. An *effect* on X is a morphism $e\colon X \to \mathsf{I}$, depicted with no output. A morphism $r\colon \mathsf{I} \to \mathsf{I}$, drawn with no input or output, is called a *scalar*.



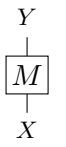
In particular the 'empty space' diagram (4.1) is the scalar $1 = \mathrm{id}_{\mathsf{I}}$.

The compositions $\circ, \otimes$ satisfy axioms which must be considered when working symbolically but are trivial in the graphical language. An example is associativity of composition $(h \circ g) \circ f = h \circ (g \circ f)$, which is automatic from simply drawing three boxes in sequence on the same wire. Similarly the rule $(f \otimes g) \circ (f' \otimes g') = (f \circ f') \otimes (g \circ g')$, displayed in the left identity below, and the 'interchange law' $(f \otimes \mathrm{id}) \circ (\mathrm{id} \otimes g) = f \otimes g = (\mathrm{id} \otimes g) \circ (f \otimes \mathrm{id})$, displayed in the right identity below, amount to letting us freely slide boxes along wires.



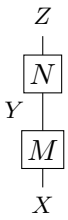
Let us now introduce our primary example category in this article.

Example 4.2.1. *In the category $\mathbf{Mat}_{\mathbb{R}^+}$ of positive valued matrices, the objects are finite sets $X, Y, \ldots$ and the morphisms $M \colon X \to Y$ are functions $M \colon X \times Y \to \mathbb{R}^+$ where $\mathbb{R}^+ := \{r \in \mathbb{R} \mid r \geq 0\}$. Equivalently such a function is given by an '$X \times Y$ matrix' with entries $M(y \mid x) := M(x, y) \in \mathbb{R}^+$ for $x \in X$, $y \in Y$.*



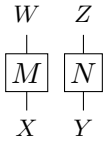
$$:: (x, y) \mapsto M(y \mid x)$$

Composition of $M \colon X \to Y$ and $N \colon Y \to Z$ is given by summation over Y:



$$:: (x, z) \mapsto \sum_{y \in Y} N(z \mid y) M(y \mid x)$$

The tensor $\otimes$ is given on objects by the Cartesian product $X \otimes Y = X \times Y$, and on morphisms by the Kronecker product, i.e. the usual tensor product of matrices:



$$:: ((x, y), (w, z)) \mapsto M(w \mid x) N(z \mid y)$$

The symmetry is simply the isomorphism $X \times Y \simeq Y \times X$. The unit object $\mathsf{I} = \{\star\}$ is the singleton set. A state of X is then equivalent to a positive function on X:

$$:: x \mapsto \omega(x)$$

where $\omega(x) := \omega(x \mid \star)$. Similarly, an effect e on X is also equivalent to a positive function on X via $e(x) := e(\star \mid x)$.

$$:: x \mapsto e(x)$$

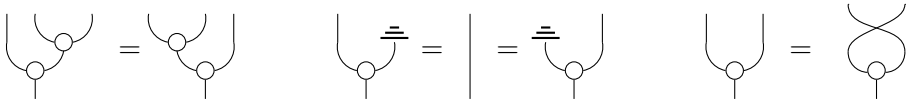
Finally, a scalar is precisely a positive real $r \in \mathbb{R}^+$.

4.2.1. Cd-categories

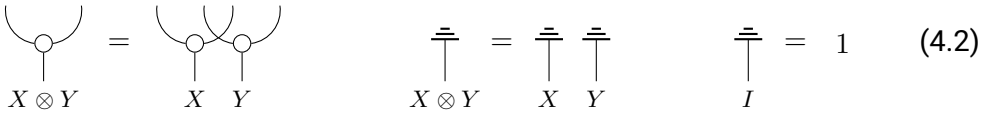
Many aspects of probability theory can be treated entirely diagrammatically, by noting that categories such as $\mathbf{Mat}_{\mathbb{R}^+}$ come with the following further structure.

Definition 4.2.2. *(Cho & Jacobs, 2019) A* cd-category *(*copy-discard category*) is a symmetric monoidal category in which each object comes with a specified pair of morphisms*

called copying *and* discarding*, respectively, which satisfy the following:*

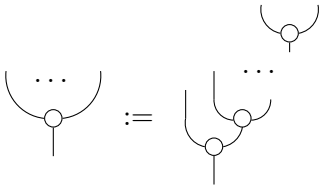


The choice of these morphisms is moreover 'natural' in that the following hold for all objects X, Y.



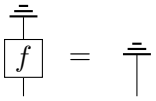
(4.2)

Thanks to these axioms for copying, we can unambiguously define a copying morphism with n output legs, for any $n \geq 1$, via:

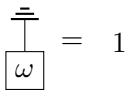


with the $n = 0$ case defined to be discarding $\top$.

The presence of discarding allows us to identify the truly 'probabilistic' processes in a cd-category. We say that a morphism f is a *channel* when it preserves discarding, as below.



In particular, we call a state ω *normalised* when the following holds. For an explanation of why this gives the usual definition, cf. Example 4.2.3 below.

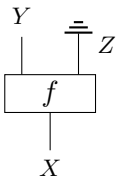


Here we will often call a normalised state ω of X a *distribution* of X, even when working in a general cd-category [2]. We also call a normalised state of $X \otimes Y$ a *joint distribution* over X, Y.

[2] This is to avoid confusion with the usual use of the term (hidden) 'state' in PP.

A cd-category in which every morphism is a channel, or equivalently ⏉ is the unique effect on any object, is called a *Markov category* (Fritz, 2020). Given any cd-category $\mathbf{C}$, its subcategory $\mathbf{C}_{\text{channel}}$ of channels always forms a Markov category.

Discarding allows us to 'ignore' certain outputs of a process. Given any morphism f from X to Y, Z, its *marginal* $X \to Y$ is the following morphism:

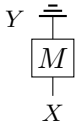


Let us see how these features describe discrete probability theory within our example category.

Example 4.2.3. $\mathbf{Mat}_{\mathbb{R}^+}$ *is a cd-category. Copying on* X *is given* $\curlyvee(y, z \mid x) = \delta_{x,y,z}$ *with value* 1 *iff* $x = y = z$ *and* 0 *otherwise. Discarding* ⏉ *on* X *is given by the function with* $x \mapsto 1$ *for all* $x \in X$*. Hence a state* ω *is normalised, i.e. forms a distribution on* X*, precisely when it forms a* probability distribution *over* X *in the usual sense, i.e. its values sum to* 1*.*

$$= \sum_{x \in X} \omega(x) \;=\; 1$$

More generally, a process $M \colon X \to Y$ *is a channel iff it forms a* probability channel*, or equivalently a* stochastic *matrix, meaning that it sends each* $x \in X$ *to a normalised distribution* $M(y \mid x)$ *over* Y*. Indeed we have that:*



$$:: \;\; x \;\mapsto\; \sum_{y} M(y \mid x)$$

Hence M *is a channel iff this effect is constant at* 1*, i.e. for all* x *we have*

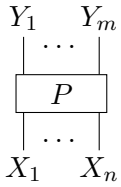
$$\sum_{y \in Y} M(y \mid x) = 1$$

In typical probability theory, such a channel is also often called a 'conditional probability distribution' $P(Y \mid X)$ *with values denoted* $P(y \mid x) := P(Y = y \mid X = x)$ *for* $x \in X$*,* $y \in Y$*. The subcategory of channels in* $\mathbf{Mat}_{\mathbb{R}^+}$ *is the Markov category* $\mathbf{FStoch}$ *of finite Stochastic matrices.*

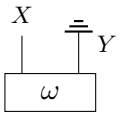
Let us see how a few features of probability theory appear in diagrams. Firstly, for any X, Y*, a distribution* ω *on* $X \otimes Y$ *corresponds to a joint distribution over* X, Y *(left-hand below). In particular given a pair of distributions* ϕ, σ *over* X, Y*, the distribution* $\phi \otimes \sigma$ *corresponds to the resulting product distribution over* $X \times Y$*, with* X *and* Y *independent*

from each-other (right-hand below).

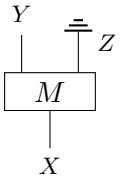
A general channel as below represents a probability channel $P(Y_1, \ldots, Y_m \mid X_1, \ldots, X_n)$.



Marginalisation of any morphism corresponds to the usual notion in probability theory, given by summation over the discarded object.



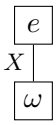
$$:: x \mapsto \sum_{y \in Y} \omega(x, y)$$



$$:: (x, y) \mapsto \sum_{z \in Z} M(y, z \mid x)$$

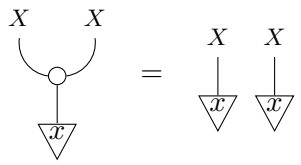
Finally we observe that for any effect $e \colon X \to \mathbb{R}^+$ *and distribution* ω *the scalar* $e \circ \omega$ *corresponds to the expectation value of the function* e *according to the probability distribution* ω.

$$\mathop{\mathbb{E}}_{x \sim \omega} e(x) \quad = \quad \sum_{x \in X} e(x)\omega(x)$$



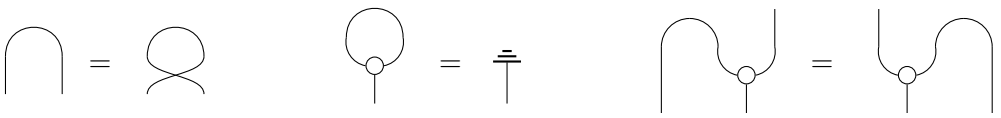
4.2.2. Sharp states and caps

The copying morphisms in a cd-category allow us to identify those states which are really 'deterministic' (Fritz, 2020). We call a state x *sharp*, and depict it with a triangle as below, when it is copied by $\curlyvee$, that is:



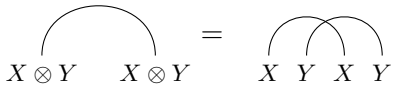
(4.3)

In many categories there is a corresponding effect for each state, playing an important role for sharp states, thanks to the following feature. We say that $\mathbf{C}$ has *caps* when each object comes with a distinguished effect on $X \otimes X$ depicted $\cap$ and satisfying:



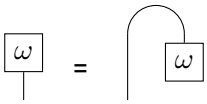
(4.4)

and such that the following holds for all objects X, Y:



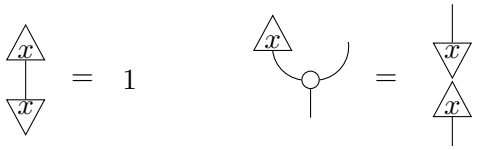
Intuitively, the cap is an effect which checks if its two input wires are 'in the same state'. The first equation in (4.4) expresses that this comparison is symmetric, and the remaining two that it is compatible with copying; for example the second says that each input when copied is equal to its copy.

Practically, caps allow us to 'turn outputs into inputs'. In particular, for each state ω we can define a corresponding effect by 'flipping ω upside-down':



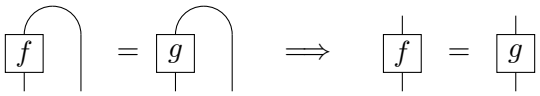
(4.5)

When $\omega = x$ is a sharp state, we call this effect *sharp* also. One may verify that it is the unique effect satisfying the following.



(4.6)

Caps are particularly useful in diagrammatic reasoning when they are *cancellative*, meaning that:



for all morphisms f, g.

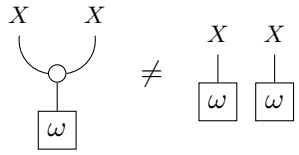
Example 4.2.4. $\mathbf{Mat}_{\mathbb{R}^+}$ *has cancellative caps. Each point $x \in X$ corresponds to a normalised sharp state on X which we again denote by x, given by the point probability distribution δ_x at X.*

$$:: \; y \mapsto \begin{cases} 1 & x = y \\ 0 & \textit{otherwise} \end{cases}$$

The corresponding effect is given by the function δ_x also. Each cap is given by $\frown(x, y) = \delta_{x,y}$. We note a useful fact that for any morphism $M \colon X \to Y$ its values $M(y \mid x)$ can be given diagrammatically as below.

$$M(y \mid x) \quad =$$

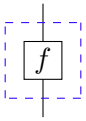
Every sharp state on X is of the above form for some $x \in X$, or else given by the zero state 0 *defined by $0(x) = 0$ for all $x \in X$. The only sharp scalars are 0 and 1. Note that a general state ω, even when normalised, is not copyable.*



Indeed the left-hand side is the distribution $(x, y) \mapsto \omega(x)\delta_{x,y}$, while the right is $(x, y) \mapsto \omega(x)\omega(y)$, which differ unless ω is zero or $\omega = \delta_x$ for some $x \in X$.

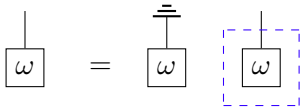
4.2.3. Normalisation

In graphical probabilistic reasoning it is also useful to be able to normalise states and processes. We say that a cd-category $\mathbf{C}$ has *normalisation* when it comes with a rule assigning each morphism $f\colon X \to Y$ a new morphism called the *normalisation* of f, depicted by drawing a dashed blue box:



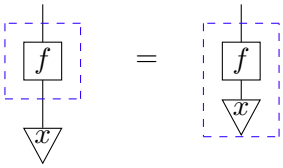
(4.7)

such that these normalisations satisfy various axioms, of which we sketch a few here. For a full definition see (Lorenz & Tull, 2023). Firstly, a general state ω is equal to a scalar multiple of its normalisation. In particular in $\mathbf{Mat}_{\mathbb{R}+}$ when the state is non-zero, this means that its normalisation is indeed normalised in our above sense, i.e. a distribution.



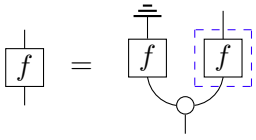
(4.8)

For a general morphism f its normalisation is given on each sharp state x by normalising $f \circ x$.



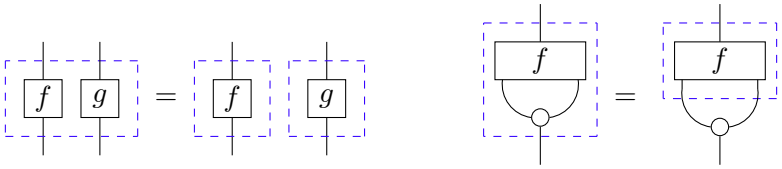
(4.9)

These two rules combine to give the following equation without explicit reference to states.



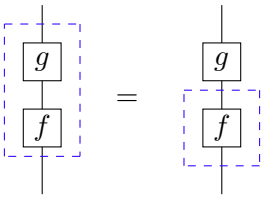
(4.10)

Note that if f we already a channel then it would be equal to its normalisation, as in this case we can passing the discarding through f and then the copy map above. In general normalisations satisfy a few graphical conditions including the following.



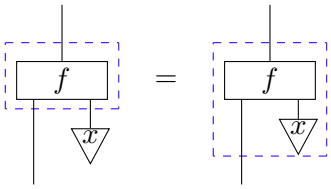
(4.11)

Further, for all morphisms f and channels g we have:



(4.12)

and for all sharp states x and morphisms f we have the following.



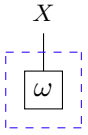
For a full account of the properties of normalisation see (Lorenz & Tull, 2023). We note that for a general morphism f, its normalisation is not necessarily a channel but only a 'partial channel' [3]. In terms of states, this is because its sends each sharp state x either to a normalised state, or else to 0 if $f \circ x = 0$. However in $\mathbf{Mat}_{\mathbb{R}^+}$ it will be a channel provided f has 'full support', so that $f \circ x$ is non-zero for all non-zero sharp states x.

Throughout the article, the following notation will be useful. For any set X and function $f \colon X \to \mathbb{R}^+$ let us write

$$\underset{x}{\mathrm{Norm}}\, f(x) := \frac{f(x)}{\sum_{x' \in X} f(x')}$$

whenever this is well-defined, i.e. the denominator is finite and non-zero.

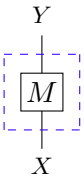
Example 4.2.5. $\mathbf{Mat}_{\mathbb{R}^+}$ *has normalisation. On each object* X *the zero state* 0*, given by* $0(x) = 0$ *for all* $x \in X$*, is defined to have normalisation* 0*. For any non-zero state* ω *we indeed have*



$$:: \; x \mapsto \underset{x}{\mathrm{Norm}}\, \omega(x)$$

[3] Such morphisms are called 'quasi-total' in (Di Lavore & Román, 2023), where morphisms satisfying (4.10) are also called 'normalisations', though are not uniquely chosen unlike our definition.

For a general morphism $M\colon X \to Y$ the normalisation is given by:



$$:: (x, y) \mapsto \begin{cases} \mathrm{Norm}_y\, M(y \mid x) & \text{if } \sum_{y \in Y} M(y \mid x) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

As a result if M has full support, so that $M(y \mid x) \neq 0$ for some y, for all x, then its normalisation is a probability channel.

4.2.4. Further cd-categories

Though we will not need them here, we note that the notion of a cd-category is much more general than $\mathbf{Mat}_{\mathbb{R}^+}$, and give a few examples for those familiar with them. The category $\mathbf{Rel}$ whose objects are sets and morphisms are relations is a cd-category, as are its subcategories $\mathbf{PFun}$ of sets and partial functions and $\mathbf{Set}$ of sets and functions, with the latter forming the channels in $\mathbf{PFun}$.

There are also many more cd-categories of a 'probabilistic' nature, see for instance (Panangaden, 1998; Cho & Jacobs, 2019; Fritz, 2020). In particular to treat general probability spaces (including 'continuous probability channels') one may work in the category $\mathrm{Kl}(G)$ of measurable spaces $X = (X, \Sigma_X)$ and Markov (sub-)kernels $f\colon X \to Y$, which send each $x \in X$ to a (sub-)probability measure $f(x)$ over Y. Roughly, this means replacing all instances of summation Σ in $\mathbf{Mat}_{\mathbb{R}^+}$ above with integration $\int$. Of particular interest in PP is the following subcategory, though we will not work with it in detail in this article.

Example 4.2.6. *(Fritz, 2020, Section 6) In the category $\mathbf{Gauss}$ the objects are spaces $X = \mathbb{R}^n$ and morphisms $M\colon X \to Y$ are Markov kernels $f\colon X \to Y$ with densities of the form $f(y \mid x) = \eta(y - Mx)$ for some fixed Gaussian noise distribution η (independent of x) and linear map $M\colon X \to Y$. This category models linear processes with Gaussian noise. More general non-linear Gaussian processes are studied in PP under the so-called 'Laplace assumption'.*

4.3. Generative Models

A central feature in PP is that each cognitive agent possesses a generative model which describes their internal beliefs about how the observations they receives arise from hidden states of the world[4]. In its simplest form, a generative model consists of a channel $c\colon S \to O$ describing how likely $c(o \mid s)$ a given observation $o \in O$ is for each hidden

[4] Note that is distinct from whatever 'true' external process produces the observations in reality, with the latter often called the 'generative process' to distinguish it from the agent's own 'generative model' (Parr et al., 2022).

state $s \in S$, along with a distribution σ over S describing prior beliefs about how likely each state is.

However, generative models typically come with further compositional structure, relating various spaces of observations and hidden states, as formalised by a *Bayesian network* (or more precisely a *causal* Bayesian network, see later discussion), a probabilistic graphical model based on a directed acyclic graph (DAG). There is in fact a close correspondence between DAGs and cd-categories, allowing us to describe and study such models entirely in terms of string diagrams. This view also leads one to consider more general 'open generative models', which may come with 'input' variables. These open models can be used to which describe the individual components of an overall generative model in the usual sense. For more details on the approach used here, see (Lorenz & Tull, 2023).

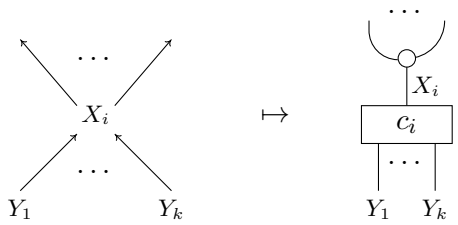
We begin by relating DAGs with the following class of string diagrams.

Definition 4.3.1. *(Lorenz & Tull, 2023) A* network diagram *is a string diagram D built from single-output boxes, copy maps and discarding:*

with labellings on the wires, such that any wires not connected by a sequence of copy maps are given distinct labels, and each label appears as an output at most once and as an input to any given box at most once.

Such diagrams are best understood by examples, which we come to shortly. Before this, we note that network diagrams are in fact equivalent to DAGs in the following sense. By an *open DAG* we mean a finite DAG G with vertices $V = \{X_1, \dots, X_n\}$, along with subsets $I, O \subseteq V$ of *input* and *output* vertices, respectively, such that each input vertex has no parents in G.

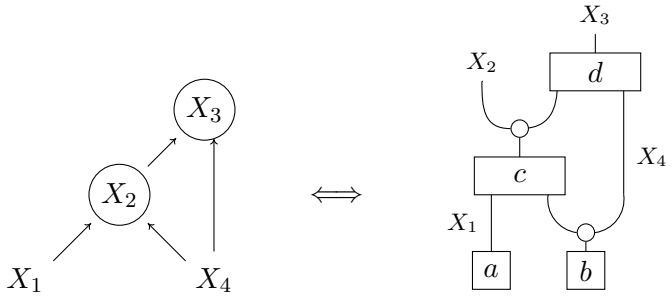
Given any open DAG G, we may construct an equivalent network diagram featuring a box c_i with output X_i for each non-input vertex X_i. The box c_i itself has an input wire for each parent of X_i in G. In the diagram we copy the output of this box and pass it to each of the children of X_i, as well as an extra time if $X_i \in O$ i.e. X_i is an output vertex of the DAG.



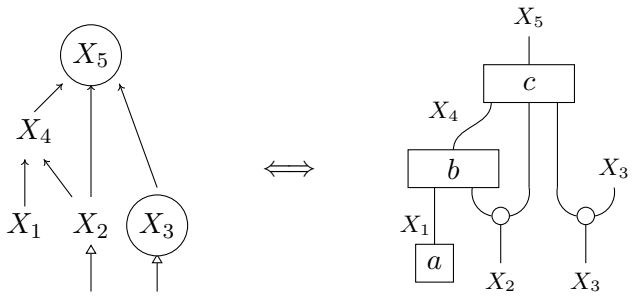
By construction, this yields a network diagram D_G from the inputs I to the outputs O of the DAG. Conversely, given any such network diagram D we define an open DAG $G_D = (G, I, O)$ with a vertex $X \in V$ for each wire X in D, and with $X \in I, O$ iff it is an input (resp. output) to the diagram.

In practice the labellings of the boxes are arbitrary, and we consider any two network diagrams equivalent when they are the same up to the equations of a cd-category and box re-labellings. Then the above yields a one-to-one correspondence between open DAGs and network diagrams (Lorenz & Tull, 2023, Sections 3,5).

Example 4.3.2. *Consider the open DAG G over $\{X_1, X_2, X_3, X_4\}$ below, with output vertices $O = \{X_2, X_3\}$ circled, and with no input vertices. The equivalent network diagram D_G is shown to the right. Note that the labels of the boxes are arbitrary.*



Example 4.3.3. *The following depicts an open DAG over $V = \{X_1, \ldots, X_5\}$ with outputs $O = \{X_3, X_5\}$ and with inputs $I = \{X_2, X_3\}$ highlighted with special incoming arrows. To the right we show the corresponding network diagram with the same inputs and outputs.*



We may now define generative models themselves, which involve specifying actual channels corresponding to the boxes in the network diagram.

An *interpretation* $[\![-]\!]$ of a network diagram D in a cd-category $\mathbf{C}$ consists of specifying an object $[\![X_i]\!]$ for each wire X_i and channel $[\![f]\!] \colon [\![X_1]\!] \otimes \cdots \otimes [\![X_k]\!] \to [\![X]\!]$ for each box f in D with inputs $X_1, \ldots, X_k$ and output X.

Definition 4.3.4. *(Lorenz & Tull, 2023) Let $\mathbf{C}$ be a cd-category. An* open generative model *in $\mathbf{C}$ is given by a network diagram D along with an interpretation $[\![-]\!]$ in $\mathbf{C}$. We call the objects corresponding to output wires* observed *and the rest* hidden*. We call such a model* closed *when it has no inputs.*

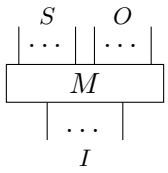
Note that an object of an open model may be both an input and output. In practice, we omit the $[\![-]\!]$ symbols and for each wire X in the network diagram of a model denote the corresponding object $[\![X]\!]$ in $\mathbf{C}$ also by X. Similarly for each box c in the diagram with output X we also write c for the corresponding channel $[\![c]\!]$.

Remark 4.3.5. *Formally, an open generative model in our sense is the same as an* open causal model *in the sense of (Lorenz & Tull, 2023); that is, both have the same mathematical definition. However a 'generative model' typically refers to a causal model with the extra interpretation of being possessed by a cognitive agent.*

Indeed, though not often stressed in the literature, a typical generative model in PP may be seen as a causal *Bayesian network, i.e. a* causal model *in the sense of Pearl (Pearl, 2009). This means that the probability channels which constitute the network do not represent arbitrary relationships but in fact (beliefs about) causal ones, such as how observations are caused by (rather than merely correlated with) hidden states of the world. For more discussion see Section 4.10.*

Given any open generative model $\mathbb{M}$ we obtain an overall channel from its inputs to its outputs by composing the channels of the model, i.e. viewing the (interpreted) network diagram as a single channel in $\mathbf{C}$. Often it is useful to also consider the following related channel.

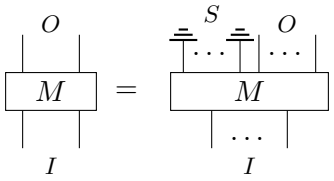
Definition 4.3.6. *Let $\mathbb{M} = (D, [\![-]\!])$ be an open generative model in $\mathbf{C}$ with inputs $I$and outputs O. Let S denote the non-input hidden (non-output) objects of the model. The* total channel M *of the model is the channel from I to S, O:*



(4.13)

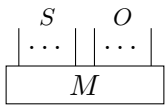
given by interpreting the network diagram D' in which we modify D by adding an extra copy morphism to each object in S, to make it an output.

Conversely, the usual channel from inputs to outputs is then simply the marginal over S:



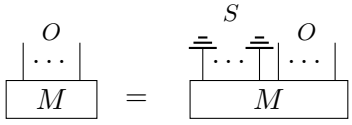
(4.14)

In particular for a closed generative model, with no inputs, we call the total channel the *total distribution* of the model. It is a joint distribution over the hidden objects S and observed objects O:



(4.15)

with the original distribution over the observed objects as its marginal.



(4.16)

Let us now consider generative models in our main example category.

Example 4.3.7. *A closed generative model $\mathbb{M}$ in $\mathbf{Mat}_{\mathbb{R}^+}$ is precisely a* Causal Bayesian Network (CBN). *This consists of specifying:*

- *a finite DAG $G$with a subset $O \subseteq V$ of 'observed' vertices and the remaining $S = V \setminus O$ being 'hidden';*
- *for each vertex X_i an associated variable with a finite set of values also denoted X_i, and a mechanism c_i given by a probability channel with density:*

$$P(X_i \mid \mathsf{Pa}(X_i)) \tag{4.17}$$

The term 'causal' refers to the fact each such mechanism has a causal interpretation.

Indeed, as we have seen, such a DAG G with outputs O is equivalent to a network diagram D with no inputs. Specifying an interpretation of D is then the same as choosing the sets X_i of values and channels (4.17) *for each box in the diagram. A CBN defines a joint distribution[5] over all the variables $V = \{X_1, \ldots, X_n\}$ with density*

$$P(V) := \prod_{i=1}^{n} P(X_i \mid \mathsf{Pa}(X_i)) \tag{4.18}$$

which is precisely (4.15). *The output distribution of the CBN is given by the marginal $P(O)$ over only the observed variables, corresponding to* (4.16).

Example 4.3.8. *An open generative model $\mathbb{M}$ in $\mathbf{Mat}_{\mathbb{R}^+}$ is an 'open CBN', where now for the input variables no channel* (4.17) *is specified. This induces via* (4.13) *the total channel*

$$P(S, O \mid I)$$

from the inputs to the non-input hidden variables S and output variables O, which here we would denote (the entries of) by $M(s, o \mid i)$. Its marginal $P(O \mid I)$ on the observed variables O yields the channel $M(o \mid i)$ from (4.14).

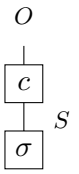
In short, a (closed) generative model in $\mathbf{Mat}_{\mathbb{R}^+}$ specifies the internal structure of an output distribution $P(O)$ in terms of further variables and channels (4.17), while an open generative model similarly specifies the internal structure of a channel $P(O \mid I)$ from inputs I to outputs O.

For the remainder of this section we will describe some of the common forms of (open) generative models which appear in PP.

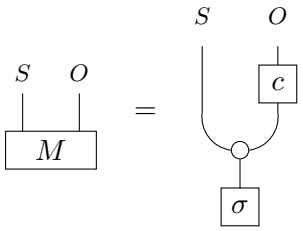
[5] Often a Bayesian network is instead defined as a distribution $P(V)$ satisfying the *Markov condition* (4.18) in terms of its conditionals. Since these conditionals may not be unique, and the channels c_i are an important component of the model, we instead include the latter explicitly; for more discussion see (Lorenz & Tull, 2023).

4.3.1. Simple generative models

By a *generative model* $S \to O$ we mean a generative model $\mathbb{M}$ with network diagram:

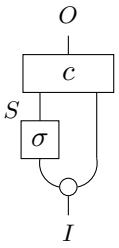


Thus $\mathbb{M}$ consists of objects S, O with O observed and S hidden, a channel $c\colon S \to O$, called the *likelihood*, and a distribution σ on S, called the *prior*. As alluded to earlier, we call S the *hidden states* and O the *observations* of the model. The total distribution of the model is given by



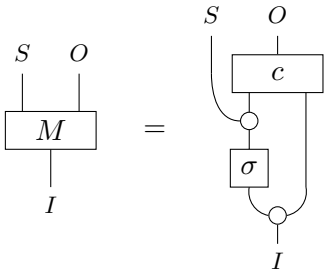
(4.19)

More generally, we can consider an *open* variant of such a generative model which now comes with a hidden object I of *inputs*, with the following network diagram:



(4.20)

Hence both the prior and likelihood now take an additional I input. The total channel is given by



(4.21)

Intuitively, such an open model $\mathbb{M}$ consists of specifying a particular generative model $S \to O$ for each input in I.

Example 4.3.9. *A generative model $S \to O$ in $\mathbf{Mat}_{\mathbb{R}^+}$ consists of a finite set S of hidden states, O of observations, a likelihood channel $c(o \mid s)$ and prior distribution $\sigma(s)$. Often would often write $c(o \mid s)$ as simply $P(o \mid s)$ and $\sigma(s)$ as $P(s)$. We interpret $c(o \mid s)$ as the probability of observing o when in the hidden state s. Then the total state* (4.19) *is the joint distribution over S, O given by*

$$M(s, o) = c(o \mid s)\sigma(s)$$

and typically simply denoted $P(s, o)$. As the notation suggests $P(o \mid s)$ is the conditional and $P(s)$ the marginal of the joint distribution $P(s, o)$.

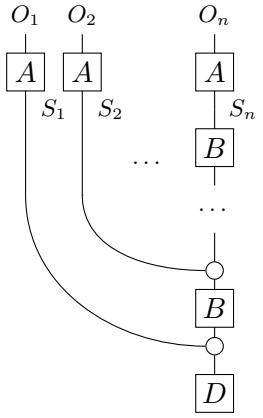
When the generative model is open as in (4.20) *it now comes with a finite set I of input values, with likelihood $c(o \mid i, s)$ and prior $\sigma(s \mid i)$. The total channel* (4.21) *is then given by*

$$M(s, o \mid i) = c(o \mid i, s)\sigma(s \mid i)$$

Thus for each input i we obtain a generative model $\mathbb{M}(i)$ of the form $S \to O$ and an induced joint distribution $M(i)$ over S, O.

4.3.2. Discrete time models

For a given $n \in \mathbb{N}$, a *discrete time generative model* is a closed generative model $\mathbb{M}$ of the form



Thus it consists of observed objects $O_1, \ldots, O_n$, hidden objects $S_1, \ldots, S_n$, a *prior* distribution D over S_1, *observation* channels $\{A \colon S_t \to O_t\}_{t=1}^{n}$ and *transition* channels $\{B \colon S_t \to S_{t+1}\}_{t=1}^{n-1}$. Typically we mean that there are fixed objects S, O such that $S_j = S, O_j = O$ for all j, and similarly as our notation suggests all the A and B channels for each time step are taken to be identical.

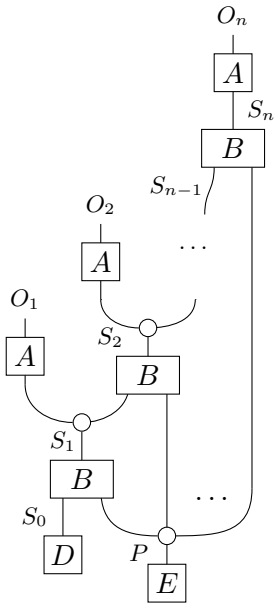
We interpret the model $\mathbb{M}$ as describing the evolution of a system over discrete time steps from $t = 1, \ldots, n$. The system begins in its initial state with prior distribution D and then evolves over each time step according to the transition channels B. Independently, we observe the system at each time t via the channel A to produce an observation in O_t.

Example 4.3.10. *A discrete time generative model in* $\mathbf{Mat}_{\mathbb{R}^+}$ *is also called a* Hidden Markov Model *or* partially observable Markov decision process (POMDP) *(Parr et al., 2022).*

4.3.3. Policy models

We now introduce an explicit ingredient whereby the agent can model its own *actions*. As in reinforcement learning (Tschantz et al., 2020), a choice of actions or behaviour is called a *policy*. In a discrete time setting, a policy can be thought of as determining likely sequences of actions over the time steps, which in turn influence the evolution of the states over time.

An n*-time step model with policies* is a generative model $\mathbb{M}$ of the form:



Thus it now includes a hidden object P of *policies* which forms an input to each transition channel B from S_t, P to S_{t+1}, for $t \leq n-1$. The model also comes with a prior distribution E over P, which are called the *habits* of the system. Note that here the policy the system is undertaking is considered hidden.

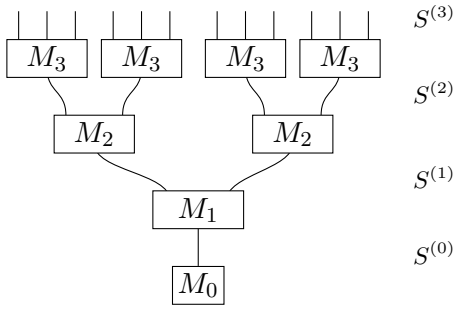
Again we typically take $S_j = S$ and $O_j = O$ for some fixed objects S, O, with all channels A identical and all B channels identical.

Example 4.3.11. *Models of this form, within* $\mathbf{Mat}_{\mathbb{R}^+}$*, are the central examples used in the active inference tutorial (Smith et al., 2022) and book (Parr et al., 2022).*

4.3.4. Hierarchical models

Central to much of PP is the study of *hierarchical* generative models (De Vries & Friston, 2017; Parr et al., 2022), which have a natural graphical description. These are generative

models given by composing various open generative models in layers, where the outputs of the open models in one layer match the inputs of the models in the next layer, such as in the example below.



(4.22)

Here it is understood that each box M_j represents an open generative model, which we may decompose further in terms of its own network diagram with inputs and outputs as shown. The right-hand labels indicate that the input wire to M_1 has type $S^{(0)}$, the output wires from M_1 both have type $S^{(1)}$ etc [6].

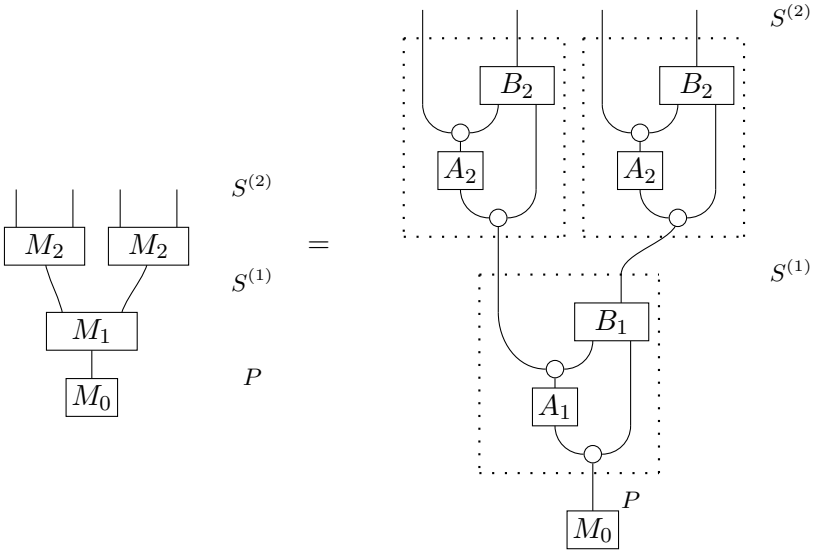
We interpret the inputs to each (box within a) layer as a 'control' signal from the layer below. Note that because we read diagrams bottom to top, the layers further down the diagram are in fact those usually referred to as more 'high-level' or 'higher' in the hierarchy.

The structure of the model tells us that the 'high-level' features cause the generation of the 'lower-level' features. For example $S^{(0)}$ could describe an overall action policy while the $S^{(3)}$ control more fine-grained motor actions. Another common example explored in (De Vries & Friston, 2017) is where the output wires from each box denote individual time steps. In this case time runs faster in the lower-level layers (higher in the diagram). For example in the diagram above six time steps occur in layer $S^{(3)}$ for every time step in layer $S^{(1)}$.

Plugging in the network diagrams for each open model corresponding to M_j yields a composite network diagram for the whole hierarchical model. For example in the following hierarchical model, the network diagrams for M_1 and M_2 are shown in the highlighted

[6] It is also common to introduce a labelling convention for the wires such as $S^{(0,1,2)}$ where the indices represent wire numbers in each layer as we read up the diagram. However this quickly becomes unwieldy, and in most cases the graphical description of the network is the most convenient.

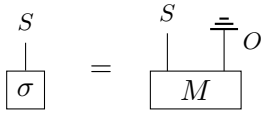
boxes below and compose to yield the diagram on the right-hand side.



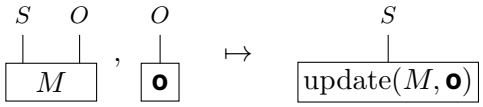
Much of the PP literature concerns such hierarchical models and the passing of these 'top-down predictions' (the flow of information up the diagram in this case) are adjusted by 'bottom-up errors' passed back down the model. The latter takes place when a model is *updated*, which we address next.

4.4. Updating Models

Consider an agent with be a simple generative model $\mathbb{M}$ of the form $S \to O$ as in Section 4.3.1. Recall that this induces a joint distribution M over S, O as in (4.15), whose marginal on S is the prior σ describing 'beliefs' about how likely each state in S is to occur.



Now suppose the agent receives an observation, which in general may be 'soft', given by an distribution **o** over O. The agent would like to *update* these beliefs to obtain a new *posterior* distribution over S, describing how likely each $s \in S$ now is given the observation.



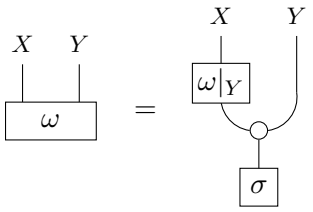
How then should the agent update the marginal on S to yield this posterior? For a general soft observation with distribution **o** over O there are at least two distinct but natural ways to carry out Bayesian-style updating, as pointed out by Jacobs in (Jacobs, 2019), which

we describe in this section. When the observation **o** is sharp, however, corresponding to a single element $o \in O$, there is a canonical way to carry out this belief updating, usually simply referred to as *Bayesian* updating, which we introduce first.

4.4.1. Sharp Updating

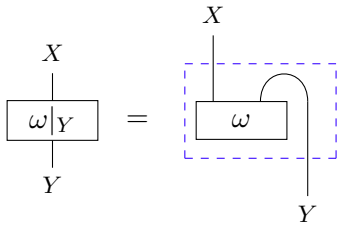
Let us begin by describing updates with respect to a sharp observation, given by (a point distribution at) an element $o \in O$. Such Bayesian updating is closely related to the notion of *conditional* probabilities, which have a nice characterisation in cd-categories. Here we follow the approach to conditioning from (Lorenz & Tull, 2023), building on earlier treatments (Coecke & Spekkens, 2012; Cho & Jacobs, 2019; Fritz, 2020); see also (Di Lavore & Román, 2023).

Definition 4.4.1. *Let $\mathbf{C}$ be a cd-category, and ω a joint distribution over X, Y. Then* a conditional *of ω by Y is a morphism $\omega|_Y \colon Y \to X$ such that the following holds:*



(4.23)

where σ is the marginal of ω on Y. If $\mathbf{C}$ has normalisation and cancellative caps, we define the (minimal) conditional *to be the morphism:*



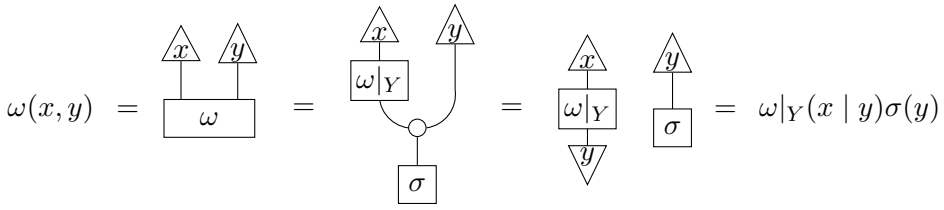
Each minimal conditional is indeed a conditional as shown in the Appendix of (Lorenz & Tull, 2023). As we saw for normalisations, a conditional is only a partial channel in general, being a channel only when ω has 'full support'.

Example 4.4.2. *In $\mathbf{Mat}_{\mathbb{R}^+}$ the minimal conditional $\omega|_Y$ is given by*

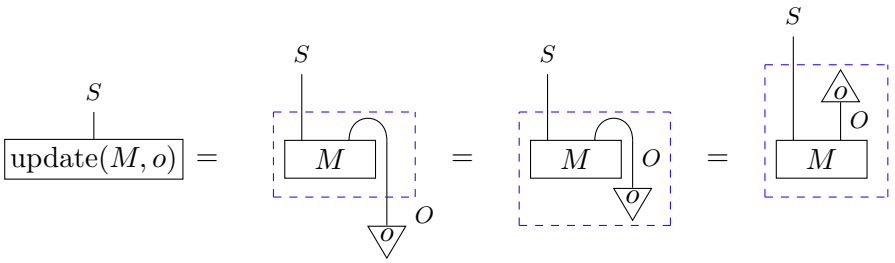
$$\omega|_Y(x \mid y) := \operatorname*{Norm}_x \omega(x, y) = \frac{\omega(x, y)}{\sum_{x'} \omega(x', y)}$$

whenever the sum in the denominator is non-zero, and $\omega|_Y(x \mid y) = 0$ for all x otherwise. Thus when ω is normalised with density denoted $P(X, Y)$ this is the usual conditional $P(X \mid Y)$. The condition (4.23) *amounts to the usual 'chain rule' $P(x, y) = P(x \mid y)P(y)$*

for the probability distribution $P(x, y) = \omega(x, y)$, since $\sigma(y)$ is the marginal $P(y)$ and we have:



For a generative model $\mathbb{M}$ of the form $S \to O$ with joint distribution M over S, O we call the minimal conditional $M|_O \colon O \to S$ the *Bayesian inverse* of the model. It specifies how to update beliefs about S for each specific sharp observation $o \in O$. Explicitly, given a sharp distribution $\mathbf{o} = \delta_o$ over O for some $o \in O$ the updated beliefs are given by the posterior:



(4.24)

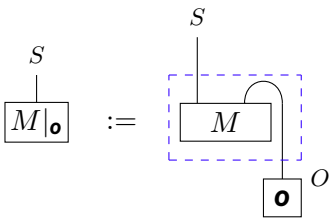
Example 4.4.3. *In $\mathbf{Mat}_{\mathbb{R}^+}$, for a sharp observation δ_o for some $o \in O$, the posterior is the distribution over S given by the usual Bayesian update:*

$$M(s \mid o) = \frac{M(s, o)}{\sum_{s'} M(s', o)} \tag{4.25}$$

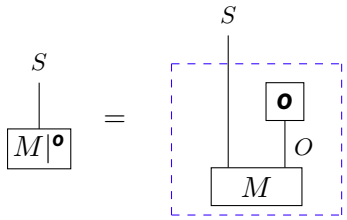
4.4.2. Pearl and Jeffrey Updating

There are two distinct ways to generalise updating to the case of a soft observation given by a distribution $\mathbf{o}$ over O, described in (Jacobs, 2019). Diagrammatically these correspond to generalising from either the former or latter diagrams in (4.24). For more on both forms of updating in cd-categories see also the treatment by Di Lavore and Román (Di Lavore & Román, 2023).

Definition 4.4.4. *Let $\mathbf{C}$ be a cd-category with normalisation and cancellative caps, and M a joint distribution over S, O. Given a distribution $\mathbf{o}$ over O, the* Jeffrey update *denoted M_J or $M|_{\mathbf{o}}$ is given by the composite $M|_O \circ \mathbf{o}$, i.e.:*



whenever this is normalised, and more generally is given by the normalisation of the above state. The Pearl update *denoted M_P or $M|^{\mathbf{o}}$ is instead given by the normalisation:*



recalling that the effect **o** *is given by composing* **o** *with a cap as in* (4.5).

Example 4.4.5. *For a generative model $\mathbb{M}$ from S to O in $\mathbf{Mat}_{\mathbb{R}^+}$, with joint distribution M over S, O, the Jeffrey update is given by*

$$M_J(s) = \mathop{\mathbb{E}}_{o\sim\mathbf{o}} \mathop{\mathrm{Norm}}_{s} M(s,o) = \sum_o \frac{M(s,o)\mathbf{o}(o)}{\sum_{s'} M(s',o)} \tag{4.26}$$

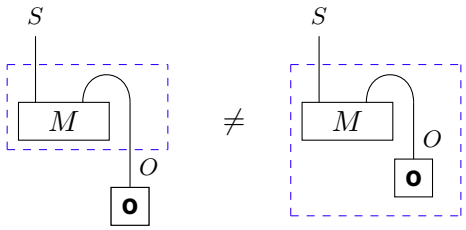
while the Pearl update is

$$M_P(s) = \mathop{\mathrm{Norm}}_{s} \mathop{\mathbb{E}}_{o\sim\mathbf{o}} M(s,o) = \frac{\sum_o M(s,o)\mathbf{o}(o)}{\sum_{s',o'} M(s',o')\mathbf{o}(o')} \tag{4.27}$$

The distinction between both update procedures is not always considered in the literature since for sharp observations they coincide with the usual Bayesian update. Indeed the following is immediate from (4.24).

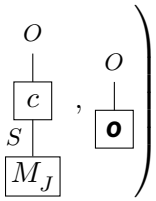
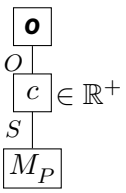
Lemma 4.4.6. *Let $\mathbf{C}$ be a cd-category with normalisation and cancellative caps. Then for each sharp state o on O the updates coincide: $M_J = M_P = M|_O \circ o$.*

In contrast, for a general observation **o** the two updates differ in the way they apply normalisation, amounting to whether one normalises with respect to (or separately from) the observation itself.



The Jeffrey update simply composes the observation **o** with the Bayesian inverse (partial) channel $M|_O$. If $M|_O$ is only a partial channel the result may not be normalised (such as when **o** falls outside the support), in which case the update is then further normalised. The Pearl update instead involves a single normalisation, taking place after composing with the observation, so that **o** is inside the normalisation box.

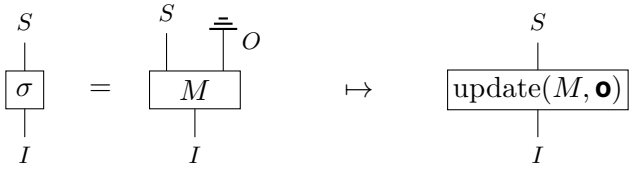
Remark 4.4.7. *Jacobs has compared the two forms of updating within $\mathbf{Mat}_{\mathbb{R}^+}$ in detail, noting that their inferences can differ considerably, but that both can be considered rational notions of updating (Jacobs, 2019). One difference between the updates is that by definition Jeffrey updating forms a probability channel in O (whenever $M|_O$ is a channel, i.e. M has full support over O). In contrast, the normalisation over* **o** *in Pearl updating means that it does not form a channel in O. The two update procedures can also be characterised by the following respective properties. For a generative model $\mathbb{M}$ over S, O with likelihood c, Jeffrey updating minimises the KL-divergence between* **o** *and the marginal on O of the updated model in which we replace the prior with the posterior (left-hand below). Pearl updating instead has the property that it maximizes the expected value of the function* **o** *(right-hand below).*

M_J *minimises:* $D_{KL}\left(\right.$  $\left.\right)$ M_P *maximises:* 

The PP literature has mostly focused on updating with respect to sharp observations, in which the two notions coincide. It is an interesting question for the future to determine which (if either) form of updating is most natural in Bayesian models of cognition.

4.4.3. Updating Open Models

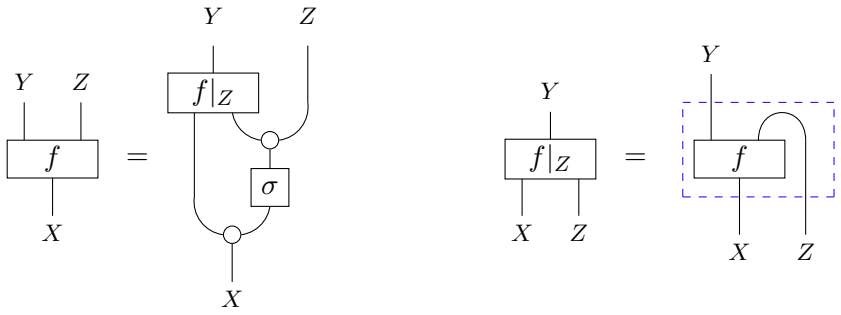
Since a typical generative model in PP is composed of various *open* generative models $\mathbb{M}$, it is also important to describe how an agent may update such open models $\mathbb{M}$, now coming with inputs I. In this case we consider the induced channel $M\colon I \to S, O$. The prior beliefs about S are now given by the marginal $\sigma\colon I \to S$, which we can think of specifying beliefs over S for each input $i \in I$. Given an observation **o** over O the agent now wishes to update this to a posterior channel of the same kind.



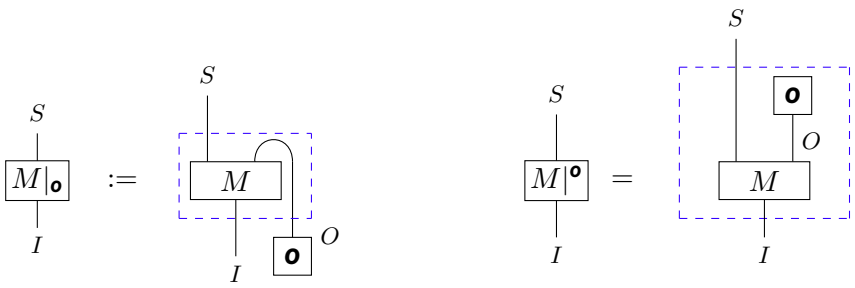
All of the treatment of updating above generalises straightforwardly to such open models, amounting to updating the corresponding closed model $M(i)$ over S, O for each input $i \in I$.

Explicitly, for any morphism $f\colon X \to Y \otimes Z$ in a cd-category, *a conditional* is any morphism $f|_Z$ satisfying the left-hand equation below, where σ is the corresponding marginal of f. In the presence of normalisation and cancellative caps, *the (minimal)*

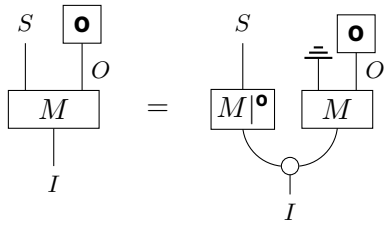
conditional is that given on the right below, as in (Lorenz & Tull, 2023).



Definition 4.4.8. *Let $M\colon I \to S, O$ be the channel induced by an open model $\mathbb{M}$, and* **o** *a distribution over O, in a cd-category with normalisation and cancellative caps. The* Jeffrey update *denoted M_J or $M|_{\mathbf{o}}$ is given by composing $M|_O$ with* **o** *as left-hand below (or more generally by its normalisation if the result is not a partial channel). The* Pearl update *denoted M_P or $M|^{\mathbf{o}}$ is instead given as on the right-hand side.*



By the defining property of normalisations (4.10) the Pearl update $M|_{\mathbf{o}}$ satisfies the following, which will be useful later.



(4.28)

Example 4.4.9. *In $\mathbf{Mat}_{\mathbb{R}^+}$ the minimal conditional of f by Z is given by*

$$f|_Z(y \mid x, z) = \operatorname*{Norm}_y f(y, z \mid x)$$

and for a probability channel $P(Y, Z \mid X)$ corresponds to the usual conditional $P(Y \mid X, Z)$. The formulae for both updates $M_J(s \mid i), M_P(s \mid i)$ are the same as (4.26)*,* (4.27) *simply replacing each $M(s, o)$ term with $M(s, o \mid i)$, i.e.*

$$M_J(s \mid i) = \mathop{\mathbb{E}}_{o \sim \mathbf{o}} \operatorname*{Norm}_s M(s, o \mid i)$$
$$M_P(s \mid i) = \operatorname*{Norm}_s \mathop{\mathbb{E}}_{o \sim \mathbf{o}} M(s, o \mid i)$$

Again both update procedures coincide with $M(s \mid o, i)$ for sharp observations $i \in I$ and all inputs $i \in I$.

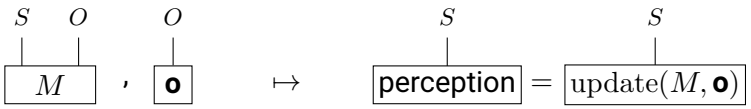
Remark 4.4.10. *Di Lavore and Román also study both forms of updating in cd-categories in which (non-chosen) conditionals exist in (Di Lavore & Román, 2023), calling them 'partial Markov categories'. There updating is defined via arbitrary (non-minimal) conditionals, meaning that $M|_O$ can be arbitrarily defined outside the support on O of M. However since this arbitrary choice can impact the result of a Jeffrey update $M|_O \circ \mathbf{o}$ when $\mathbf{o}$ is also non-zero outside this support, we instead define updating via the minimal conditional $M|_O$.*

4.5. Perception and Planning

Let us now see how the notion of updating is applied by an agent to govern its behaviour in PP. Two fundamental uses of updating are the following.

Perception Firstly, as already alluded to, we can consider the case of an agent with a generative model $\mathbb{M}$ from S to O, interpreted as accounting for observations O in terms of hidden states of the world S. For example, O may be the space of pixel-level descriptions of images while S is a compressed representational space of possible objects which the images portray.

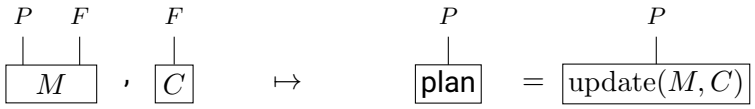
Given an observation encoded by a (soft or sharp) distribution $\mathbf{o}$ over O, the agent can update its prior over hidden states S to obtain a posterior describing how likely each hidden state is to have caused the observation. We refer to this general process of updating as *perception* and view the resulting distribution as the agent's specific perception of the observation $\mathbf{o}$. Intuitively perception takes the 'raw data' of the observation $\mathbf{o}$ and returns (a distribution over) representations S.



Intuitively, the update answers the question *'Given that I have received this observation, how likely is each possible world state?'*. In the literature this is often referred to as *inference*, in reference to Bayesian inference.

Planning A second application of updating by an agent is in *planning* its behaviours. Here an agent possesses a generative model $\mathbb{M}$ of the same formal structure but with objects labelled P, F and interpreted differently. Now P encodes the action policies, or behaviours, the agent may carry out, while F represents observations (or states) it may receive in the *future*. The model $\mathbb{M}$ includes a prior over policies which we can think of as the agent's habits or typical behaviours.

Here the agent possesses some *preferences* about which future observations (or states) are most desirable, encoded by a distribution C over F. Intuitively, the distribution will have highest density on the most desirable outcomes. The agent can then plan its actions by updating its habits with respect to these preferences:



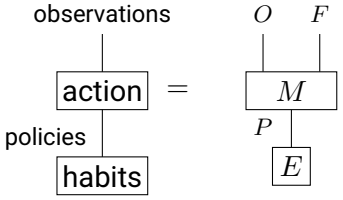
The process of deriving this distribution can intuitively be called 'planning'. We can think of this update as answering the question *'Given that I will obtain my preferences in the future, how likely is each policy to have led to this outcome?'*.

The resulting 'plan' distribution over P can be used to guide the agent's future behaviour. For example, an agent may then sample an policy to pursue from this distribution, so that the more probable policies according to the distribution are more likely to be carried out.

4.6. Exact Active Inference

Both uses of updating by an agent, planning and perception, come together in the concept of *active inference*, of which we are now able to present a fully formal diagrammatic account.

Consider an agent possessing a generative model describing how its actions, in the form of action policies P, bring about changes in its observations. These consist of both observations for the present time (and previous times) O and for future time steps F. Thus the agent has a closed generative model $\mathbb{M}$ of the following form.

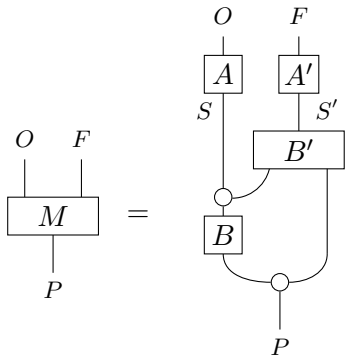


(4.29)

Here (abusing notation slightly) we denote by M also the channel from policies to observations induced by the model, and E is the prior over policies describing the agent's habits.

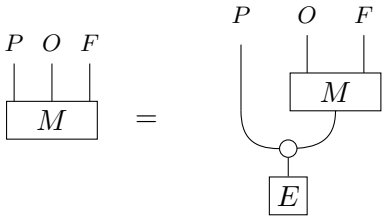
Suppose further that the agent's model explains the observations at each of these time steps through hidden states, where S denotes the hidden states in the present

time and S' in the future, so that we have:



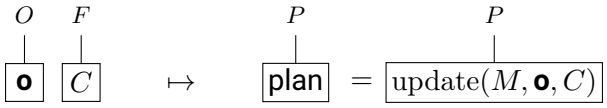
(4.30)

for 'observation' channels A, A' and 'transition' channels B, B'. The induced distribution on P, O, F is then given by:



(4.31)

The goal of active inference is then the following. The agent receives a current observation given by a distribution **o** over O, and also carries a distribution C describing its preferences for future observations F. The agent then wishes to update its prior E over policies to yield a posterior which describes its plan of action[7]:



(4.32)

Intuitively, the posterior over policies can be thought of as answering the question *'Given that I have received this observation* **o** *now, and will attain my preferences* C *in the future, which action policy am I pursuing?'*. Note that, perhaps surprisingly, the agent's own action policy is thus treated as hidden from itself, and something that it must infer.

Now, typically the objects above all decompose into further structure, as in the following example.

Example 4.6.1. *A common application of active inference is to the discrete-time models with policies given in Section 4.3.3, which we may view as instances of* (4.30) *as follows.*

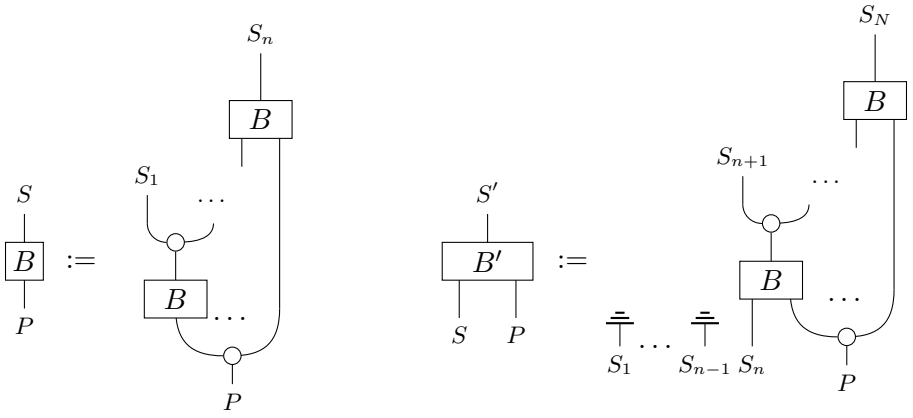
[7] Ultimately, having derived their 'plan' distribution the agent may then sample a single action policy $\pi \in P$ as in Section 4.5, and act accordingly. We imagine that via the true 'generative process' in the world (distinct from the agent's model) this leads to further observations in the future, to which the agent carries out further planning steps, and so on. Our focus is simply on a single step of how the agent derives their 'plan' from **o** and C.

Consider such a model featuring N time-steps, where $n << N$ is considered the current time, and all times m with $n \leq m \leq N$ as in the future. The spaces of 'current' hidden states and observations S, O are the products over all previous time-steps $t = 1, \ldots, n$ up to and including the current time, while the future hidden states and observations S', F take the product over all future time-steps $t = n+1, \ldots, N$.

$$S := S_1 \otimes \cdots \otimes S_n \qquad S' := S_{n+1} \otimes \cdots \otimes S_N$$
$$O := O_1 \otimes \cdots \otimes O_n \qquad F := O_{n+1} \otimes \cdots \otimes O_N$$

The observation channels in the overall model (4.30) *would then be given by:*

while the transition channels are as follows:



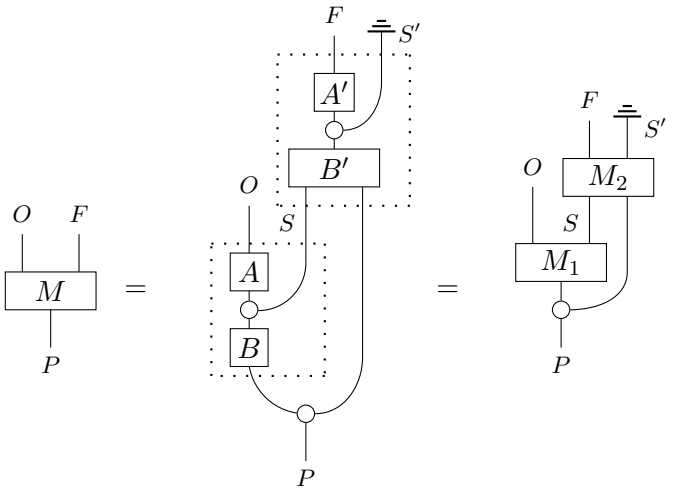
so that the composite (4.31) *yields the network diagram for the overall model for times $t = 1, \ldots, N$.*

An agent may employ various update procedures, such as those discussed in Section 4.4, to calculate its plan of action (4.32). Though both forms of updating coincide for sharp inputs, and the observations **o** in the active inference literature are typically taken to be sharp, the preferences C are often not; that is, there may be multiple desirable future observations in F. Thus Pearl and Jeffrey updating can be expected to differ.

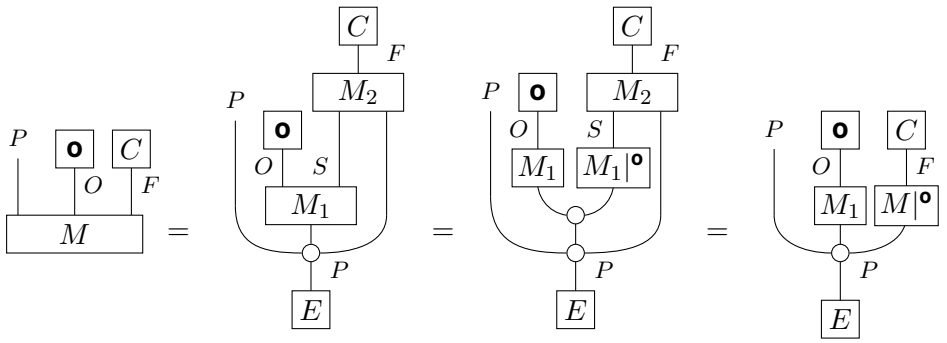
Here we will describe an exact active inference procedure based on Pearl updating, allowing both observations **o** and preferences C to be soft. We leave the exploration of Jeffrey updating in active inference for future work.

Now let us consider how the agent can in the ideal case compute its plan (4.32) via

an exact update procedure. Firstly, let us rewrite the channel in (4.30) as follows.



Here the channels M_1, M_2 are the compositions indicated by the highlighted boxes[8]. Now applying the property of Pearl updates (4.28) to M_1 we have the following:

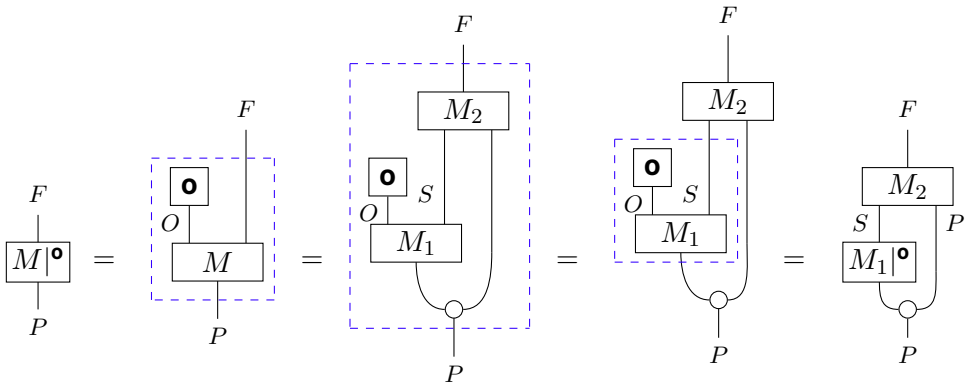


(4.33)

Here we have again denoted by M_1, M_2 their respective marginals on O, F, given by discarding S, S' respectively. In the last step we used associativity of copying and the

[8] While we could define M_2 without S' as an output, the appearance of S' will be useful later in treating approximate active inference. Note also that the dashed boxes in this case do not denote normalisation.

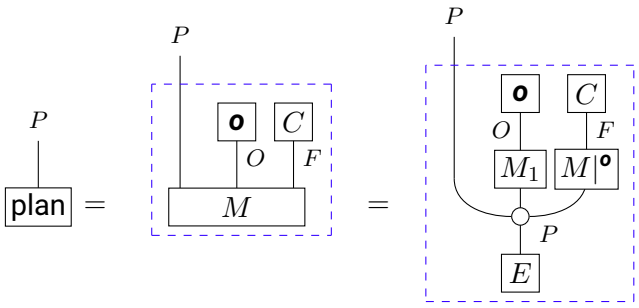
following argument:



where in the middle step we used (4.11) and (4.12) to slide the channel M_2 and copying out of the normalisation box, respectively.

Thus we obtain an exact expression for active inference.

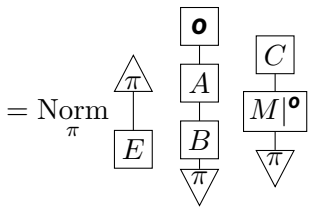
Proposition 4.6.2. *The plan over policies in Pearl-style exact active inference is given by:*



(4.34)

In $\mathbf{Mat}_{\mathbb{R}^+}$ *the plan has density over policies* $\pi \in P$ *given by:*

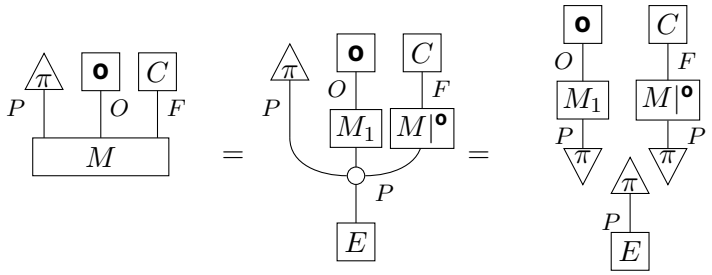
$$\mathsf{plan}(\pi) := \underset{\pi}{\mathrm{Norm}}\left(E(\pi)(\mathbf{o} \circ M_1(\pi))(C \circ M|^{\mathbf{o}}(\pi))\right) \tag{4.35}$$



(4.36)

Proof. The first equality holds by definition, so $\mathsf{plan}(\pi) = \mathrm{Norm}_\pi f(\pi)$ where f is the

density of the state in (4.33). But this is given by:



using that π is sharp, where the three right-hand scalars are precisely the terms in (4.35). The last line comes from noting that the given marginal $M_1 \colon P \to A$ is precisely $B \circ A$. □

There is only one problem with this form of active inference: the quantity (4.35) is completely intractable to calculate. Along with the normalisation in calculating $M|^{\mathbf{o}}$, calculating the terms in (4.35) would involve summation (or integration) over S, O and S', F respectively, requiring us to respectively calculate:

$$\sum_{s \in S, o \in O} \mathbf{o}(o) A(o \mid s) B(s \mid \pi) \qquad \sum_{o' \in F} C(o') M|^{\mathbf{o}}(o' \mid \pi)$$

To make the calculation of these updates tractable, an agent in active inference is understood to instead use a special form of approximation scheme, to which we now turn.

4.7. Free Energy

We have seen that for an agent to perform exact Bayesian updating is computationally intractable. In active inference, an agent instead carries out approximate updating by minimising a quantity known as *free energy* (K. Friston et al., 2006; K. Friston, 2010; Parr et al., 2022). In this section for simplicity we work concretely in the category $\mathbf{C} = \mathbf{Mat}_{\mathbb{R}^+}$, though the same notions should be similarly defined in continuous settings.

The extra mathematical ingredient[9] needed to define free energy will be the following .

Definition 4.7.1. *For any distribution σ over X and $x \in X$ we define the* surprise *as* $\mathrm{S}(\sigma)(x) := -\log \sigma(x)$. *For another distribution ω on X we define the overall* surprise *of σ relative to ω as the expectation value:*

$$\mathrm{S}\left(\boxed{\omega}, \boxed{\sigma}\right) \quad := \quad - \mathop{\mathbb{E}}_{x \sim \omega} \log \sigma(x)$$

[9] To study free energy we will move beyond a purely diagrammatic approach and make use of some probabilistic calculations, most notably to define 'surprise'. However later in Section 9 we will see how to represent surprise in diagrams (via 'log-boxes'). In future work it would be interesting to represent all of the calculations in this section using such diagrams.

The entropy $\mathrm{H}(\omega)$ *of* ω *is its self-surprise:*

$$\mathrm{H}\left(\boxed{\omega}\right) := \mathrm{S}\left(\boxed{\omega}, \boxed{\omega}\right)$$

while the Kullback-Liebler (KL) divergence $\mathrm{D}(\omega, \sigma)$ *from* σ *to* ω *is the difference between these quantities:*

$$\mathrm{D}\left(\boxed{\omega}, \boxed{\sigma}\right) := \mathrm{S}\left(\boxed{\omega}, \boxed{\sigma}\right) - \mathrm{H}\left(\boxed{\omega}\right)$$

The KL divergence is a commonly used similarity measure on distributions, with $D(\omega, \sigma) \geq 0$ and $D(\omega, \omega) = 0$ for all distributions ω, σ.

We may now define the following general notion of free energy. Throughout we consider a distribution M over S, O, which we imagine to be induced by a generative model from S to O. In this section for simplicity given any such distribution we denote its marginals on S, O and conditional channels $M|_S, M|_O$ again simply by M.

Definition 4.7.2 (Free Energy)**.** *The* Free Energy *of a distribution* Q *over* S, O *relative to* M *is defined as:*

$$\mathrm{FE}\left(\boxed{Q}^{S\ O}, \boxed{M}^{S\ O}\right) := \mathrm{S}\left(\boxed{Q}^{S\ O}, \boxed{M}^{S\ O}\right) - \mathrm{H}\left(\boxed{Q}^{S}\right) \tag{4.37}$$

Explicitly then we can re-write the free energy in the following useful form.

$$\begin{aligned}
\mathrm{FE}(Q, M) &= \mathop{\mathbb{E}}_{(s,o)\sim Q}[\log(Q(s)) - \log(M(s, o))] && (4.38)\\
&= \mathop{\mathbb{E}}_{(s,o)\sim Q}[\log(Q(s) - \log(M(s \mid o)) - \log M(o)] && (4.39)\\
&= \mathop{\mathbb{E}}_{o\sim Q} \mathrm{S}\left(\boxed{Q}^{S}_{\boxed{o}}, \boxed{M}^{S}_{\boxed{o}}\right) + \mathrm{S}\left(\boxed{Q}^{O}, \boxed{M}^{O}\right) - \mathrm{H}\left(\boxed{Q}^{S}\right) && (4.40)
\end{aligned}$$

We now turn to two specific variants of this quantity commonly considered in active inference.

4.7.1. Variational Free Energy

Suppose an agent receives an observation given by a distribution **o** over O, and wishes to perform an approximate Bayesian update of its prior beliefs about S as encoded by the marginal of M on S. It may do so by finding the distribution q over S which minimises the following quantity.

Definition 4.7.3 (Variational Free Energy)**.** *Given a distribution M over S, O and distribution* **o** *over O, the* Variational Free Energy (VFE) *of a distribution q over S is defined as:*

$$\mathrm{F}\left(\begin{array}{c} S \\ | \\ \boxed{q} \end{array}\right) := \mathrm{FE}\left(\begin{array}{cc} S & O \\ | & | \\ \boxed{q} & \boxed{\mathbf{o}} \end{array}, \begin{array}{cc} S & O \\ | & | \\ \multicolumn{2}{c}{\boxed{M}} \end{array}\right)$$

An important feature of the VFE is the following. Using the expression (4.40) and pulling the entropy term inside the expectation we see that

$$\mathrm{F}(q) \;=\; \mathop{\mathbb{E}}_{o\sim\mathbf{o}} \mathrm{D}\left(\begin{array}{c} S \\ | \\ \boxed{q} \end{array}, \begin{array}{c} S \\ | \\ \boxed{M} \\ | \\ \triangledown o \end{array}\right) + \mathrm{S}\left(\begin{array}{cc} O & O \\ | & | \\ \boxed{\mathbf{o}} & \boxed{M} \end{array}\right) \tag{4.41}$$

$$\geq\; \mathrm{D}\left(\begin{array}{c} S \\ | \\ \boxed{q} \end{array}, \begin{array}{c} S \\ | \\ \boxed{M} \\ | \\ \boxed{\mathbf{o}} \end{array}\right) + \mathrm{S}\left(\begin{array}{cc} O & O \\ | & | \\ \boxed{\mathbf{o}} & \boxed{M} \end{array}\right) \tag{4.42}$$

The inequality follows from concavity of the KL divergence and Jensen's inequality, which states that for any probability measure ω on X, measurable function $f\colon X \to \mathbb{R}$ and concave function ϕ on $\mathbb{R}$ we have

$$\mathop{\mathbb{E}}_{x\sim\omega}[\phi(f(x))] \leq \phi(\mathop{\mathbb{E}}_{x\sim\omega}[f(x)]) \tag{4.43}$$

In particular we see that the inequality (4.42) will be a strict equality whenever $\mathbf{o} = \delta_o$ is given by a sharp observation $o \in O$. In this case the minimum VFE value is given by the exact Bayesian inverse $M|_o$, with value $\mathrm{F} = -\log M(o)$. Hence for a sharp observation o, minimising the VFE minimises the KL-divergence between q and the Bayesian inverse $M|_o$, achieving approximate inversion $q \approx M|_o$. Moreover $\mathrm{F}(q)$ is an upper bound on the surprise of the observation o, and when $q \approx M|_o$ we have $\mathrm{F}(q) \approx S(o, M)$.

VFE Updating This process of minimising VFE to compute an approximate Bayesian update is central in active inference, but typically only considered for such sharp observations. Here we can now consider the more general minimisation of VFE for a soft observation given by a distribution **o**. In fact we may view this as another notion of updating for a prior over S, in addition to the two forms of updating met in Section 4.4.

Firstly, observe that in the expression (4.41) since the surprise term is constant, the distribution q which minimises $\mathrm{F}(q)$ will be that which minimises the left-hand expected KL term, which is equal to the following.

$$\mathop{\mathbb{E}}_{s\sim q}\left[\log q(s) - \mathop{\mathbb{E}}_{o\sim\mathbf{o}} \log M(s \mid o)\right]$$

This quantity will in turn be minimised when this expression over S is equal to a constant K, so that:

$$\log q(s) = \mathop{\mathbb{E}}_{o\sim \mathbf{o}} [\log M(s \mid o)] + K$$

The distribution q will be given by normalising $q(s)$ in the above expression, allowing us to ignore the constant and yielding the following notion of updating motivated by the VFE. Recall that the *softmax* of a function $f\colon X \to \mathbb{R}^+$ is defined by $\sigma(f)(x) = \mathrm{Norm}_x\, e^{f(x)}$.

Definition 4.7.4 (VFE Update)**.** *Given a joint distribution M over S, O and distribution **o** over O the* VFE update *is the posterior*

$$M_F(s) = \mathop{\mathrm{Norm}}_{s}\, e^{\mathbb{E}_{o\sim \mathbf{o}} \log M(s|o)} \tag{4.44}$$

$$= \sigma(\mathop{\mathbb{E}}_{o\sim \mathbf{o}} \log M(s \mid o)) \tag{4.45}$$

where σ denotes a softmax over S.

Similarly, for any channel M from P to S, O we define the VFE update *of its marginal $P \to S$ point-wise, by $M_F(s \mid \pi) = M(\pi)_F(s)$ for each $\pi \in P$.*

From the derivation above we see that $q = M_F$ is the distribution which minimises $\mathrm{F}(q)$. Note that, as for our other forms of updating, for a sharp observation $\mathbf{o} = \delta_o$ we have $M_F(s) = M(s \mid o)$.

To relate general VFE minimisation for a soft observation to expectation values, we will use the following form of approximation. Firstly, note that by Jensen's inequality, for any probability measure ω and real function f over X we have:

$$e^{\mathbb{E}_{x\sim\omega}[\log f(x)]} \leq \mathop{\mathbb{E}}_{x\sim\omega}[f(x)] \tag{4.46}$$

Whenever we take both sides of such an inequality to be approximately equal, let us say we are using a *log approximation*. In particular for any distributions ω, σ on X the follow holds log-approximately:

$$e^{-\mathrm{S}}\left(\boxed{\omega}, \boxed{\sigma}\right) \quad \lessapprox \quad \begin{array}{c} \boxed{\sigma} \\ | \\ X \\ | \\ \boxed{\omega} \end{array} \tag{4.47}$$

Indeed this states precisely (4.46) for the case $f(x) = \sigma(x)$. Such approximations can be used to relate free energy to exact expectation values, as follows.

Proposition 4.7.5. *Let M_F be the VFE update of M relative to a distribution **o** over O, and F its VFE value. Then the following holds log-approximately:*

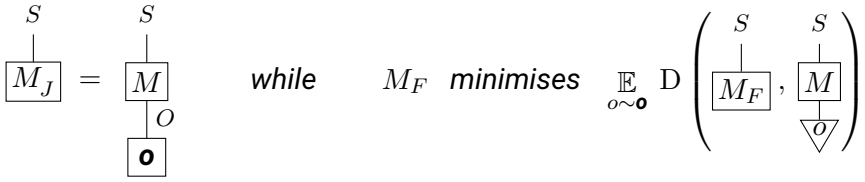
$$e^{-\mathrm{F}} \begin{array}{c} S \\ | \\ \boxed{M_F} \end{array} \approx \begin{array}{ccc} S & & \boxed{\mathbf{o}} \\ | & & | \\ \boxed{M} & & O \end{array}$$

Proof. Define $f(s) := e^{\mathbb{E}_{o\sim \mathbf{o}} \log M(s|o)}$ and the normalisation constant $K = \sum_s f(s)$, so that $KM_F(s) = f(s)$. Then we have:

$$\begin{aligned} F &= \mathrm{S}(\mathbf{o}, M) + \mathop{\mathbb{E}}_{s\sim q}[\log M_F(s) - \mathop{\mathbb{E}}_{o\sim \mathbf{o}} \log M(s \mid o)] \\ &= \mathrm{S}(\mathbf{o}, M) - \log K \\ e^{-F} M_F(s) &= e^{-\mathrm{S}(\mathbf{o},M)} K M_F(s) \\ &= e^{-\mathbb{E}_{o\sim \mathbf{o}}[\log M(o) + \log M(s|o)]} \\ &= e^{-\mathbb{E}_{o\sim \mathbf{o}}[\log M(s,o)]} \approx \mathop{\mathbb{E}}_{o\sim \mathbf{o}} M(s,o) \end{aligned} \tag{4.48}$$

where in the last step we used a log-approximation. □

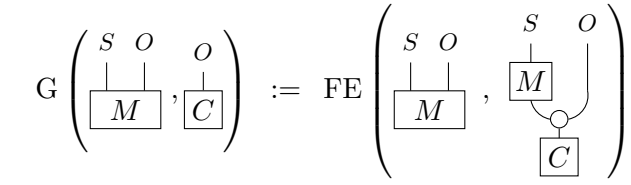
Remark 4.7.6. *Compare the formula for VFE update to the Jeffrey and Pearl updates* (4.26)*,* (4.27)*. While the Jeffrey update composes the conditional $O \to S$ with **o** exactly, the VFE update instead minimises the expected KL below.*



4.7.2. Expected Free Energy

A second form of free energy employed in active inference is used by an agent with a model featuring a space O describing observations in the future. It then has a distribution C over O modelling preferences for these future observations. Rather than updating its beliefs about future states, the agent simply want to assess how well the marginal of the model on O will fit these preferences, via the following approximation.

Definition 4.7.7. *Given a distribution M over S, O and distribution C over O, the* Expected Free Energy (EFE) *is defined as*



(4.49)

The EFE compares the given model M to the right hand generative model which perfectly attains the preferences, via its marginal C over O, whilst making use of the same inverse channel $O \to S$. Writing the EFE explicitly, and then rewriting in terms of the

typically more readily computable channel $S \to O$, we have

$$\begin{aligned}\mathrm{G}(M,C) &= \mathop{\mathbb{E}}_{(s,o)\sim M}[\log(M(s)) - \log(M(s \mid o))] - \mathop{\mathbb{E}}_{o\sim M}[\log C(o)] \\ &= \mathop{\mathbb{E}}_{\substack{s\sim M \\ o\sim M\circ s}}[-\log(M(o \mid s)] + \mathop{\mathbb{E}}_{o\sim M}[\log M(o) - \log C(o)] \\ &= \mathop{\mathbb{E}}_{s\sim M}\left[H\left(\begin{array}{c}O\\|\\\boxed{M}\\|\\\boxed{s}\end{array}\right)\right] + D\left(\begin{array}{c}O\\|\\\boxed{M}\end{array}, \begin{array}{c}O\\|\\\boxed{C}\end{array}\right)\end{aligned}$$

The final line expresses the EFE in terms of a right-hand *risk* term, which assesses how well the predicted state over O matches the preferences C, and a left-hand *uncertainty* term given by the expected entropy in the observations. Thus minimising EFE requires both matching preferences and reducing uncertainty. For more interpretations of EFE see (Parr et al., 2022).

Now using Jensen's inequality and the concavity of entropy, one may show that for any distribution ω and channel c we always have:

$$\mathop{\mathbb{E}}_{x\sim\omega} H\left(\begin{array}{c}\\|\\\boxed{c}\\|\\\boxed{x}\end{array}\right) \leq H\left(\begin{array}{c}\\|\\\boxed{c}\\|\\\boxed{\omega}\end{array}\right)$$

Hence the EFE is bounded above by the surprise of the preferences:

$$\mathrm{G}\left(\begin{array}{cc}S & O\\|&|\\\multicolumn{2}{c}{\boxed{M}}\end{array}, \begin{array}{c}O\\|\\\boxed{C}\end{array}\right) \;\leq\; H\left(\begin{array}{c}O\\|\\\boxed{M}\end{array}\right) \;+\; D\left(\begin{array}{c}O\\|\\\boxed{M}\end{array}, \begin{array}{c}O\\|\\\boxed{C}\end{array}\right) \;=\; \mathrm{S}\left(\begin{array}{c}O\\|\\\boxed{M}\end{array}, \begin{array}{c}O\\|\\\boxed{C}\end{array}\right) \tag{4.50}$$

Thus minimising the EFE results in reducing the surprise of the preferences, making them more likely to be obtained according to the model. Taking the inequality to be an approximation and applying exponentials to both side along with a log-approximation then gives the following.

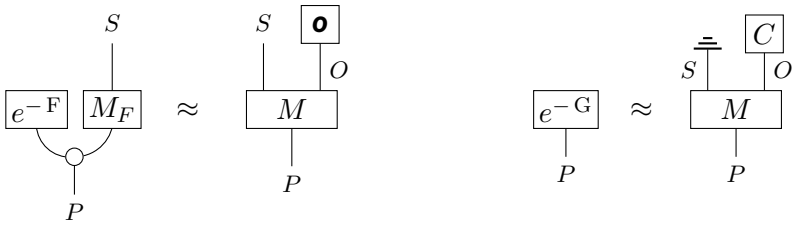
Proposition 4.7.8. *The EFE is bounded above and approximately equal to the expectation value:*

$$e^{-\mathrm{G}}\left(\begin{array}{cc}S & O\\|&|\\\multicolumn{2}{c}{\boxed{M}}\end{array}, \begin{array}{c}O\\|\\\boxed{C}\end{array}\right) \quad \lessapprox \quad \begin{array}{c}\boxed{C}\\|\\O\\|\\\boxed{M}\end{array}$$

4.7.3. Free Energy in Active Inference

We conclude this section by noting two uses of free energy in approximate active inference, treated in the next section. For these we now consider a channel M from P to S, O, typically induced by an open model. For each $\pi \in P$ this specifies a joint distribution $M(\pi)$ over S, O, to which we may apply free energy calculations.

Corollary 4.7.9. *Let* **o** *and* C *be distributions over* O*. Let* $M_F \colon P \to S$ *be the VFE update of* M *by* **o***, and for each* $\pi \in P$ *set* $\mathrm{F}(\pi) := \mathrm{F}(M(\pi)_F)$ *to the corresponding VFE value. Similarly for each* $\pi \in P$ *let* $\mathrm{G}(\pi) = \mathrm{G}(M(\pi), C)$*. Then we have the following approximations:*

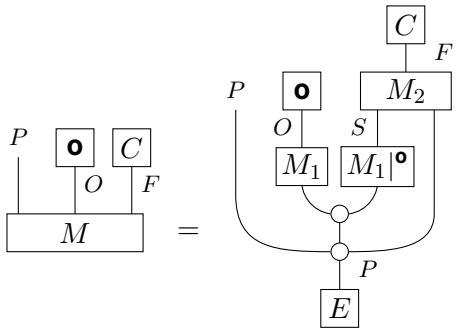


In the above the effect $e^{-\mathrm{F}}$ is given by $\pi \mapsto e^{-\mathrm{F}(\pi)}$, for $\pi \in P$, and $e^{-\mathrm{G}}$ is defined similarly.

Proof. For the first approximation, plugging in a (sharp state given by) an element $\pi \in P$ to both sides shows that this is equivalent to Proposition 4.7.5 holding for each joint distribution $M(\pi)$ over S, O with respect to the observation **o**. For the second approximation, apply Proposition 4.7.8 to the joint distribution $M(\pi)$ over S, O for each $\pi \in P$. □

4.8. Active Inference via Free Energy

Let us now return to the situation of an agent carrying out active inference as in Section 4.6. As before the agent's generative model $\mathbb{M}$ in (4.29) consists of its habits E over policies P and a channel M from P to current and future observations O, F, factoring via current and future hidden states S, S'. Given its observation **o** and future preferences C it can now use free energy to give a viable approximation of its updated plan of behaviour from Proposition 4.6.2, proceeding in two steps. We saw already in (4.33) that:



'Perception' step In the first step, the agent approximately updates the part of the model pertaining to the current time, M_1, in light of the observation **o**. For each policy π it computes a distribution $q(\pi)$ with (approximately) minimal VFE $\mathrm{F}(q(\pi))$, thus obtaining

a channel $q\colon P \to S$ which approximates the VFE update of M_1 by **o**. For each $\pi \in P$ denote the corresponding VFE value by $\mathrm{F}(\pi)$. Explicitly:

$$\mathrm{F}(\pi) = \mathrm{F}\,() := \mathrm{FE}\left(\begin{array}{c} S \quad O \\ \boxed{q(\pi)}\ \boxed{\mathbf{o}} \end{array}, \begin{array}{c} S \quad O \\ \boxed{M_1} \\ P \\ \pi \end{array}\right)$$

'Prediction' step In the second step, the agent uses this approximation channel q, to obtain a channel M_q which approximates the model over future states and observations, defined as follows:

$$\begin{array}{c} S' \quad F \\ \boxed{M_q} \\ P \end{array} = \begin{array}{c} S' \quad F \\ \boxed{M_2} \\ S \quad\quad P \\ \boxed{q} \\ \bigcirc \\ P \end{array}$$

For each policy π this induces a distribution $M_q(\pi)$ over S', F, for which the agent can compute the EFE with respect to the preferences:

$$\mathrm{G}(\pi) := \mathrm{G}\left(\begin{array}{c} S' \quad F \\ \boxed{M_q} \\ \pi \end{array}, \begin{array}{c} F \\ \boxed{C} \end{array}\right) \tag{4.51}$$

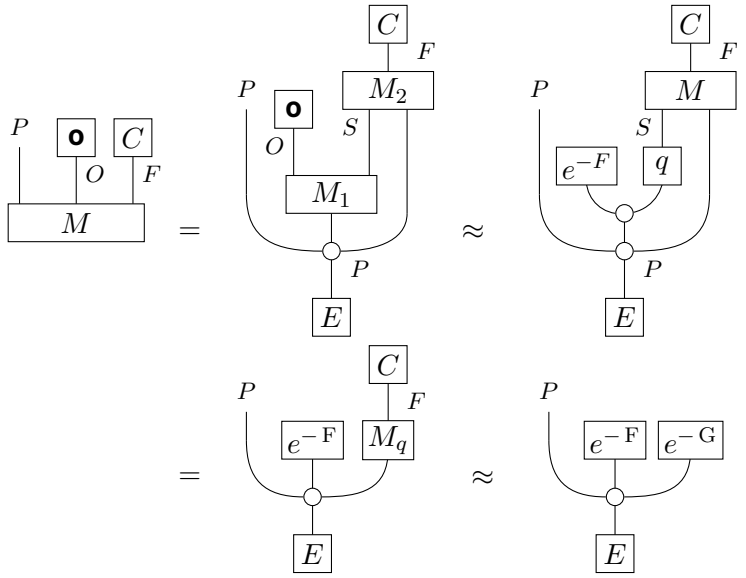
Using these free energy quantities, the agent may carry out approximate active inference. The following formula is central in the active inference literature.

Theorem 4.8.1. *The agent can carry out approximate active inference given observation* ***o*** *and preferences C by setting its plan to have density*

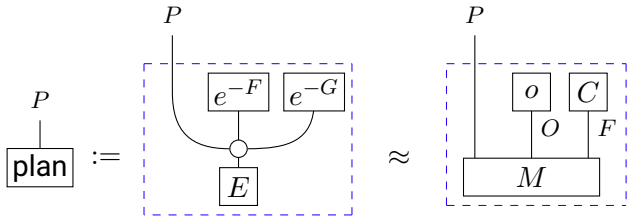
$$\mathsf{plan}(\pi) := \sigma(\log E(\pi) - F(\pi) - G(\pi)) \tag{4.52}$$

where σ denotes a softmax over $\pi \in P$.

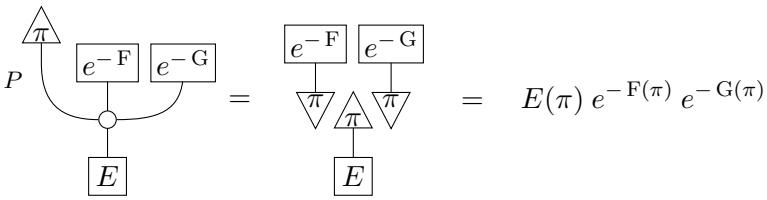
Proof. We have:



where we used Corollary 4.7.9 in both approximation steps. Thus defining our plan as on the left-hand below yields an approximate update:



But finally, note that the left-hand distribution is precisely given by (4.52). Indeed for each $\pi \in P$, corresponding to a sharp effect on P, we have:



Hence the normalisation of the above is precisely the softmax expression (4.52). □

This formula for active inference via free energy, though frequently used, is usually only justified in a fairly heuristic manner (Parr et al., 2022). Previous accounts rely on the less clear notion of treating EFE as a 'prior' to updating[10]. Here we have instead seen how the expression can be derived from a direct diagrammatic argument, directly from the structure of the generative model.

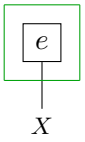
[10] Despite the fact EFE is not straightforwardly a component of the generative model, and requires inference over present states S to be calculated first, rather than prior to them

4.9. Compositionality of Free Energy

A crucial aspect of active inference is the idea that an agent can be understood to minimise free energy at all levels, so that it may be seen to globally minimise free energy in its generative model by minimising free energy within each component.

To formalise this idea we must first introduce a notion of free energy for open models. For this we will make use of the following graphical notation for the surprise.

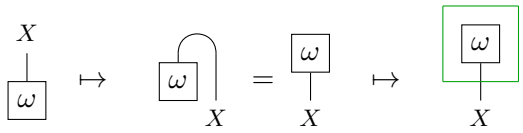
Definition 4.9.1. *Given any effect e on X in $\mathbf{Mat}_{\mathbb{R}^+}$, corresponding to a function $e\colon X \to \mathbb{R}^+$, we denote by*



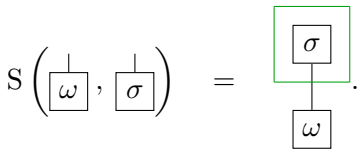
the function $-\log e(x)\colon X \to (-\infty, \infty]$.

Remark 4.9.2. *Note that a log-box is no longer an effect within $\mathbf{Mat}_{\mathbb{R}^+}$, since when $e(x) = 0$ we will have $-\log e(x) = \infty$. Here we will interpret any diagram involving log-boxes with inputs $X_1, \ldots, X_n$ and no outputs as a (formula specifying a) function $X_1 \times \cdots \times X_n \to (-\infty, \infty]$. Composing boxes in the diagram amounts to summation over wires, as for $\mathbf{Mat}_{\mathbb{R}^+}$. Given two such diagrams D_1, D_2 we write $D_1 + D_2$ for the function given by their point-wise sum as functions. In future it would be interesting to explore a formal categorical semantics for log-boxes.*

In particular we can apply a log box to any distribution ω in $\mathbf{Mat}_{\mathbb{R}^+}$ by first turning it into an effect, yielding the surprise $S(\omega)(x) = -\log \omega(x)$.



Similarly for a pair of distributions ω, σ we have

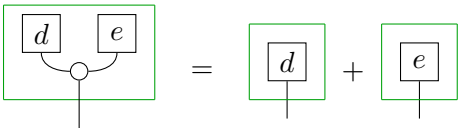


(4.53)

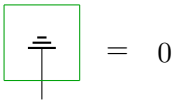
From the properties of the logarithm, one may verify that log-boxes then satisfy the following compositional properties.

Lemma 4.9.3. *For all effects d, e and sharp states x the following hold.*

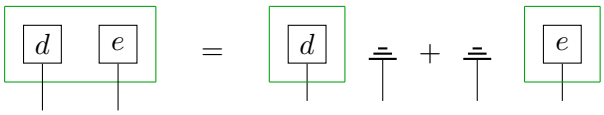
1.



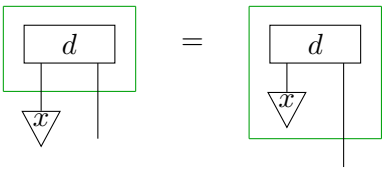
2.



3.



4.

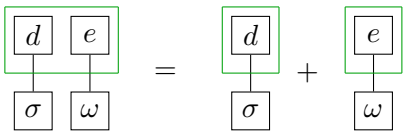


Proof. Plugging inputs x, y into each equation they reduce to the following respective properties of the logarithm. (1): $\log(d(x)e(x)) = \log(d(x)) + \log(e(x))$. (2): $\log(1) = 0$. (3): $\log(d(x)e(y)) = \log d(x) + \log e(y)$. (4) holds by definition, since both diagrams are given by $y \mapsto \log d(x, y)$. □

The following properties then follow from diagrammatic reasoning, using the relation between caps and copying.

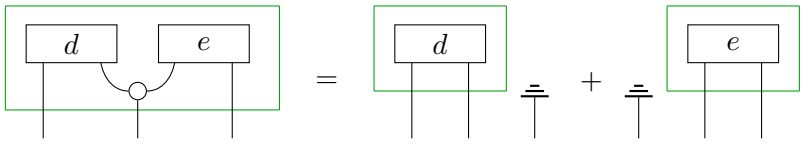
Proposition 4.9.4.

1. *For all effects d, e and normalised states σ, ω:*

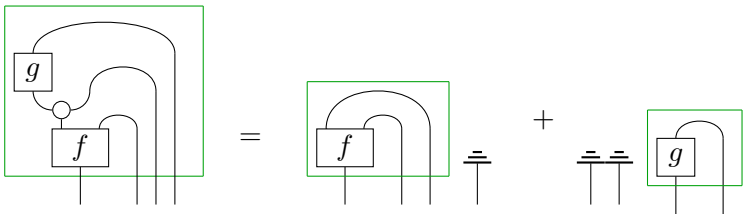


   In particular, entropy is additive across parallel composition: $H(\sigma \otimes \omega) = H(\sigma) + H(\omega)$.

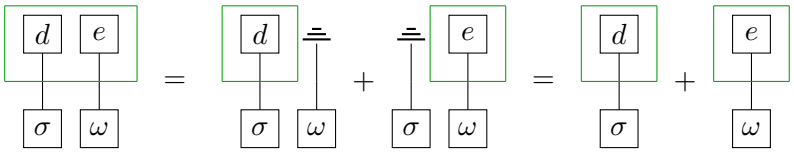
2. *For all effects d, e:*



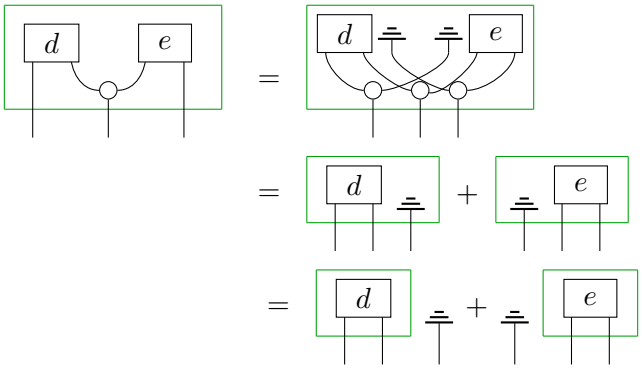
3. *For all morphisms f, g:*



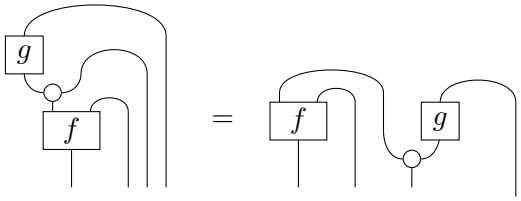
Proof. (1) follows from Lemma 4.9.3 (3) since:



(2) follows from Lemma 4.9.3 (1), 2 and 3 since:

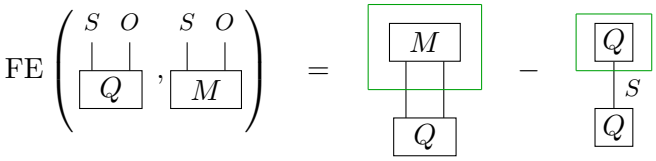


(3) is a special case of (2) where we define the effects d, e by composing f, g with caps on their output, respectively, since using the relation between caps and copying we see that:

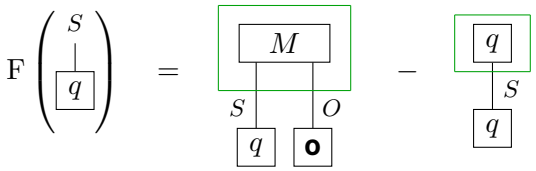


□

Now, by construction, for any joint distributions M, Q over S, O the free energy is given by:

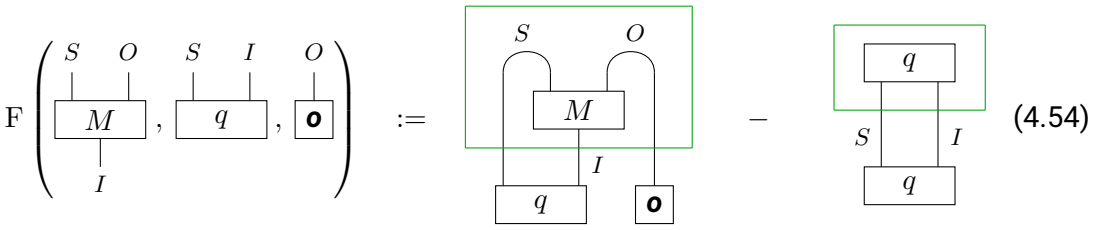


Hence for any joint distribution M over S, O and distributions q, **o** over S, O respectively, the VFE is given by:



We can use this to define a generalisation of VFE for (the channels induced by) open generative models.

Definition 4.9.5 (Open VFE)**.** *Given a channel $M\colon I \to S, O$, distribution q over I, S and distribution* ***o*** *over O, we define the* open Variational Free Energy *as*

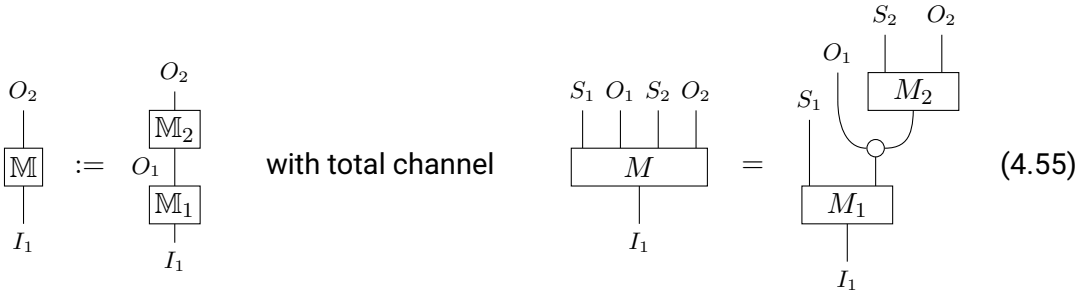


(4.54)

In the special case where I is trivial, the open VFE coincides with the usual VFE.

We can use the compositional properties of log boxes to show that this form of free energy is compositional, in an appropriate sense. First consider the following two ways in which we may compose open models.

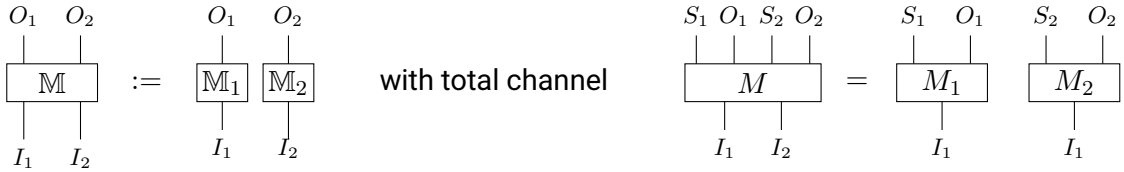
Consider a pair of open models $\mathbb{M}_1, \mathbb{M}_2$ with inputs I_1, I_2, outputs O_1, O_2 and hidden states S_1, S_2, respectively, such that $O_1 = I_2$. We can compose these in sequence into a single open model $\mathbb{M}$ from I_1 to O_2, with S_1, S_2, O_1 as its hidden states, with induced total channel M from I_1 to the remaining variables given as below.



(4.55)

Formally, the left-hand diagram is a composition in the category of open causal models; see (Lorenz & Tull, 2023, Sec. 5).

We can also compose open models in parallel. Given two open models $\mathbb{M}_i$ with inputs I_i, outputs O_i and hidden states S_i, for $i = 1, 2$ we can define an open model $\mathbb{M}$ with both sets of inputs, outputs and hidden states with induced induced total channel M from I_1, I_2 to the remaining variables given as below.

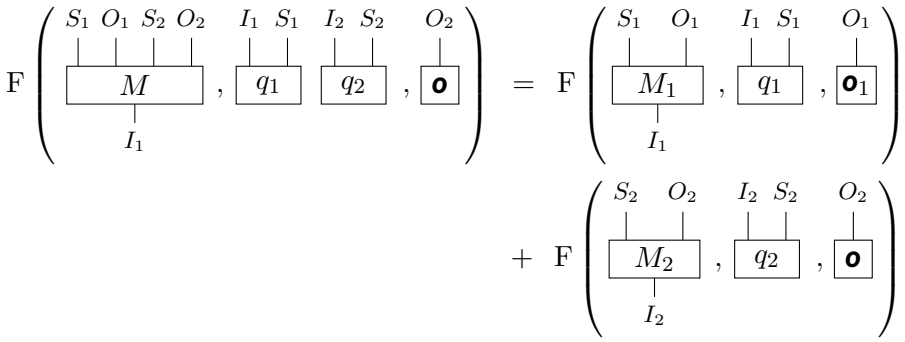


(4.56)

For each of these forms of composition of open models, we wish to establish that free energy is compositional in that the VFE of the open model $\mathbb{M}$ is determined from the

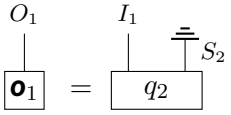
VFE of its constituents. This ensures that locally minimising VFE (within each of subcomponent) can achieve global VFE minimisation also.

Theorem 4.9.6. *For the sequential composite model $\mathbb{M}$ in* (4.55) *with $O_1 = I_2$, with total channel M, and any distribution* **o** *over O_2 and distributions q_1, q_2, the following holds:*



(4.57)

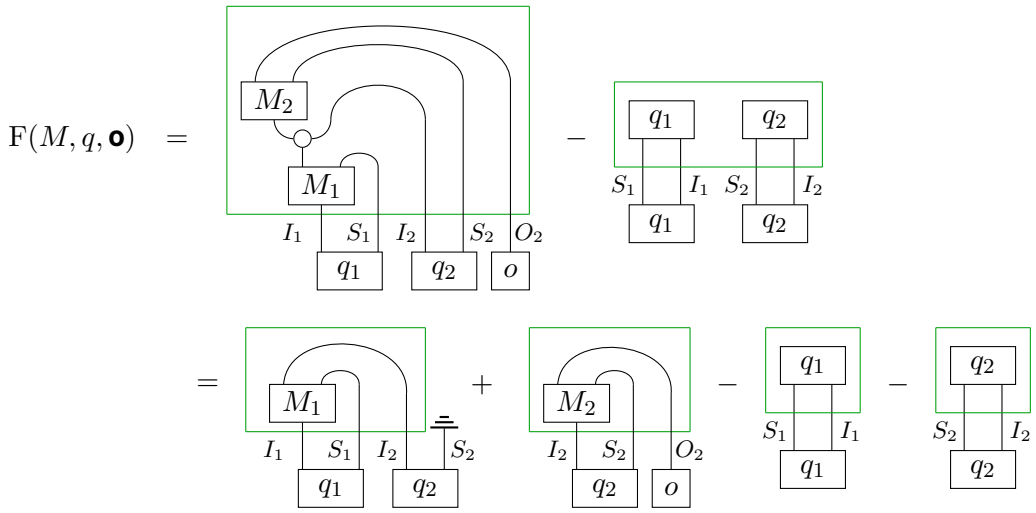
where



(4.58)

Intuitively, (4.58) expresses the way in which the beliefs about inputs I_2 in M_2 are passed down to the model $\mathbb{M}_1$ as observations in $O_1 = I_2$.

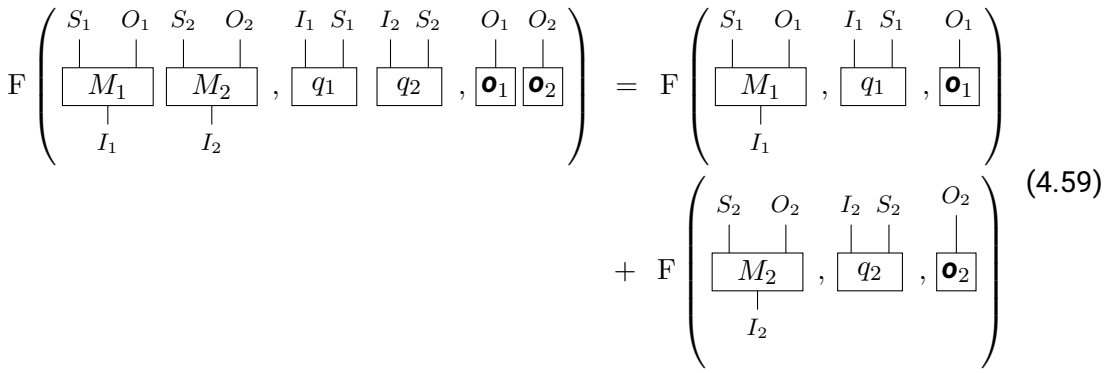
Proof. After rearranging some wires, we have that



where for the first two terms we apply Proposition 4.9.4 with $f = M_1$ and $g = M_2$, along with the fact that **o** and q_1 are normalised, and the second two terms are from Proposition 4.9.4 (1). But this is precisely the right-hand side of (4.57). □

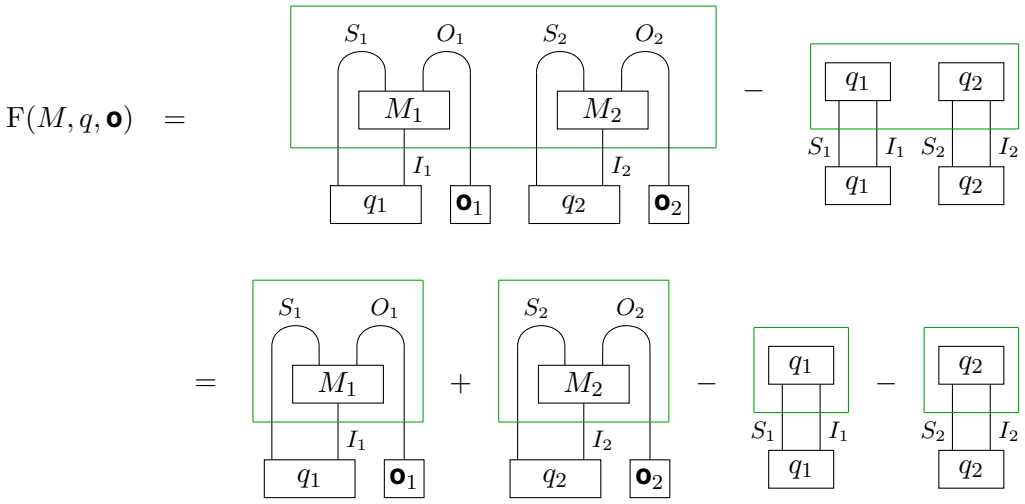
Next let us turn to the parallel composite of open models.

Theorem 4.9.7. *For any channels and distributions $M_i, q_i, \mathbf{o}_i$ for $i = 1, 2$, the following holds.*



(4.59)

Proof. The left-hand side is given by



which is precisely the right-hand side, where we applied Proposition 4.9.4 (1). □

The above results tell us that an agent with an overall generative model may minimise VFE by minimising VFE locally within each sub-model, an important property underlying the application of the free energy to all levels of a system.

4.10. Outlook

In this article we have aimed to give a concise formulation of active inference in terms of string diagrams interpreted in a cd-category $\mathbf{C}$, focusing on the case of finite discrete systems as described by $\mathbf{C} = \mathbf{Mat}_{\mathbb{R}^+}$. In particular we were able to derive the formula for approximate active inference via free energy minimisation purely from the high-level structure of a generative model undertaking active inference, and derived a compositionality property for free energy.

However these are just the first steps towards a fully compositional account of intelligent behaviour according to predictive processing, and there are many directions for future work.

Message passing So far we only studied active inference at a high-level, saying that an agent must, for each observation, arrive at an updated distribution q by free energy minimisation, without discussing how this is to be carried out. In PP this minimisation is normally achieved via so-called 'message passing' algorithms, such as 'variational' and 'marginal' message passing (Parr et al., 2019). These are defined on undirected graphical models described by Forney factor graphs, induced by a generative model. In future it would be interesting to include a categorical account of message passing within our framework, to complete our description of active inference.

Continuous settings Another technical matter would be to extend the treatment of PP beyond the finite case to further cd-categories describing continuous settings, such as a suitable category of Gaussian probabilistic processes, which are widely employed in PP under the 'Laplace assumption'. One issue is in extending our treatment of minimal conditionals to such continuous settings, where they are not as straightforwardly defined.

Causal reasoning We have here pointed out that an generative model may be seen precisely as a causal model (Pearl, 2009). In future it would be interesting to explore how an agent may carry out causal reasoning on its model using concepts from the causal model framework such as 'interventions', as treated graphically in (Lorenz & Tull, 2023), and how such reasoning relates to active inference.

Approximations The treatment of active inference via free energy in Section 4.8 relied on applying various approximation steps from Section 4.7 to parts of the diagram. Certainly more could be done to set bounds on how well these approximations hold, including how they extend from part of a diagram to the whole generative model.

Updating within PP The categorical perspective led us to naturally consider soft observations (given by distributions) rather than the usual sharp ones (given by points), which come with distinct notions of Jeffrey and Pearl updating (Section 4.4), as well our new notion of VFE updating (Section 4.7). While we were able to describe active inference via the latter two forms of updating, it would be interesting to compare against Jeffrey updating and establish which form of exact updating is most naturally considered (and approximated) in PP. That is, given that both forms of updating have different goals (Jacobs, 2019), which one (if either) is approximately carried out by the brain? This question was also raised in (Di Lavore & Román, 2023).

We note that Pearl updating can be more generally defined with respect to any effect (see e.g. (Di Lavore & Román, 2023)), i.e. any (not necessarily normalised) function. There is disagreement between active inference and reinforcement learning (RL) in whether an agent's preferences should, rather than as a distribution as in active inference, be simply modelled by a function $C\colon F \to \mathbb{R}^+$ assigning a 'value' in $\mathbb{R}^+$ to each possible future observation, i.e. as an effect C on F (K. J. Friston et al., 2009; Tschantz et al., 2020). In this case Pearl updating may be the most natural to treat planning. In contrast, Jeffrey updating may be most fitting for perception, with an observation **o** naturally encoded as a distribution i.e. a 'fuzzy point' in O.

Consciousness in PP Various proposals have been put forward for how PP and active inference can be related to consciousness. Continuing from previous work from two of the authors on IIT (Kleiner & Tull, 2021; Tull & Kleiner, 2021), in future we hope to account for these proposals within our graphical account of active inference.

Categorical modifications of PP Beyond simply recasting previous results in PP categorically, in future one may also study what new insights the compositional perspective may bring to PP and active inference, and to connect the work to ongoing research within categorical cybernetics (Smithe, 2021b; Capucci et al., 2021) and more broadly to the research programme of compositional intelligence.

Part II.

On Experiments

5. Falsification and Consciousness

Johannes Kleiner, Erik Hoel[1]

5.1. Introduction

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

At this point in the field's development falsification has become relevant. In general, scientific theories should strive to make testable predictions (Popper, 1959). In the search for a scientific theory of consciousness, falsifiability must be considered explicitly as it is commonly assumed that consciousness itself cannot be directly observed, instead it can only be inferred based off of report or behavior.

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

[1]Published as: Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, 2021(1), niab001. (Kleiner & Hoel, 2021)

"avalanche" or "ignition" of widespread neural activity created by an interconnected but dispersed network of neurons with long-range connections (Sergent & Dehaene, 2004).

Another avenue of research strives to derive a theory of consciousness from analysis of phenomenal experience. The most promising example thereof is Integrated Information Theory (Tononi, 2004, 2008; Oizumi et al., 2014). Historically, Integrated Information Theory is the first well-formalized theory of consciousness. It was the first (and arguably may still be the lone) theory that makes precise quantitative predictions about both the contents and level of consciousness (Tononi, 2004). Specifically, the theory takes the form of a function, the input of which is data derived from some physical system's internal observables, while the output of this function are predictions about the contents of consciousness (represented mathematically as an element of an experience space) and the level of consciousness (represented by a scalar value Φ).

Both Global Neuronal Workspace (GNW) and Integrated Information Theory (IIT) have gained widespread popularity, sparked general interest in consciousness, and have led to dozens if not hundreds of new empirical studies (Massimini et al., 2005; Del Cul, Baillet, & Dehaene, 2007; Dehaene & Changeux, 2011; Gosseries, Di, Laureys, & Boly, 2014; Wenzel et al., 2019). Indeed, there are already significant resources being spent attempting to falsify either GNW or IIT in the form of a global effort pre-registering predictions from the two theories so that testing can be conducted in controlled circumstances by researchers across the world (Ball., 2019; Reardon, 2019). We therefore often refer to both GNW and IIT as exemplar theories within consciousness research and show how our results apply to both. However, our results and reasoning apply to most contemporary theories, e.g. (Lau & Rosenthal, 2011; Chang et al., 2020), particularly in their ideal forms. Note that we refer to both "theories" of consciousness and also "models" of consciousness, and use these interchangeably (Seth, 2007).

Due to IIT's level of formalization as a theory, it has triggered the most in-depth responses, expansions, and criticisms (Cerullo, 2015; Bayne, 2018; P. A. Mediano, Seth, & Barrett, 2019; Kleiner & Tull, 2021) since well-formalized theories are much easier to criticize than non-formalized theories. Recently one criticism levied against IIT was based on how the theory predicts feedfoward neural networks have zero Φ and recurrent neural networks have non-zero Φ. Since a given recurrent neural network can be "unfolded" into a feedfoward one while preserving its output function, this has been argued to render IIT outside the realm of science (Doerig et al., 2019). Replies have criticised the assumptions which underlie the derivation of this argument (Kleiner, 2020a; Tsuchiya et al., 2020).

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

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

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

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

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

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

Our results show that if independence of prediction and inference holds true, as in

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

5.2. Formal Description of Testing Theories

Here we provide a formal framework for experimentally testing a particular class of theories of consciousness. The class we consider makes *predictions* about the *conscious experience* of *physical systems* based on *observations or measurements*. This class describes many contemporary theories, including leading theories such as Integrated Information Theory (Oizumi et al., 2014), Global Neuronal Workspace Theory (Dehaene & Changeux, 2004), Predictive Processing (when applied to account for conscious experience (Dolkega & Dewhurst, 2020; A. Clark, 2019; Seth, 2014; Hobson & Friston, 2014; Hohwy, 2012)) or Higher Order Thought Theory (Rosenthal, 2002). These theories may be motivated in different ways, or contain different formal structures, such as for example the ones of category theory (Tsuchiya, Taguchi, & Saigo, 2016). In some cases, contemporary theories in this class may lack the specificity to actually make precise predictions in their current form. Therefore, the formalisms we introduce may sometimes describe a more advanced form of a theory, one that can actually make predictions.

In the following section, we introduce the necessary terms to define how to falsify this class of theories: how measurement of a physical system's observables results in datasets (Section 5.2.1), how a theory makes use of those datasets to offer predictions about consciousness (Section 5.2.2), how an experimenter makes inferences about a physical system's experiences (Section 5.2.3), and finally how falsification of a theory occurs when there is a mismatch between a theory's prediction and an experimenter's inference (Section 5.2.4). In Section 5.2.5 we give a summary of the introduced terms. In subsequent sections we explore the consequences of this setup, such as how all contemporary theories are already falsified if the data used by inferences and predictions are independent, and also how theories are unfalsifiable if this is changed to a strict form of dependency.

5.2.1. Experiments

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

the set of long-range cortical connections that make up the global workspace of the brain, with p being the activity of that global workspace at that time.

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

$$obs : P \twoheadrightarrow \mathcal{O} \,, \tag{5.1}$$

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

Note that obs describes the experiment, the choice of observables, and all conditions during an experiment that generates the dataset o necessary to apply the theory, which may differ from theory to theory, such as interventions in the case of IIT. In all realistic cases, the correspondence obs is likely quite complicated since it describes the whole experimental setup. For our argument it simply suffices that this mapping exists, even if it is not known in detail.

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

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

5.2.2. Predictions

A theory of consciousness makes predictions about the experience of some physical system in some configuration, state, or dynamics, p, based on some dataset o. To this end, a theory carries within its definition a set or space E whose elements correspond to various different *conscious experiences* a system could have. The interpretation of this set varies from theory to theory, ranging from descriptions of the level of conscious experience in early versions of IIT, descriptions of the level and content of conscious experience in contemporary IIT (Kleiner & Tull, 2021), or the description only of whether

a presented stimuli is experienced in GNW or HOT. We sometimes refer to elements e of E simply as *experiences*.

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

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

Shown in Figure 5.2.1 are the full set of terms needed to formally define how most contemporary theories of consciousness make predictions about experience. So far, what we have said is very general. Indeed, the force and generalizability of our argument comes from the fact that we do not have to define $pred$ explicitly for the various models we consider. It suffices that it exists, in some form or the other, for the models under consideration.

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

5.2.3. Inferences

As discussed in the previous section, a theory is unfalsifiable given just predictions alone, and so $pred$ must be compared to something else. Ideally this would be the actual conscious experience of the system under investigation. However, as noted previously, the class of theories we focus on here assumes that experience itself is not part of the observables. For this reason, the experience of a system must be inferred separately from a theory's prediction to create a basis of comparison. Most commonly, such inferences are based on *reports*. For instance, an inference might be based on an

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

Figure 5.2.1.: We assume that an experimental setup apt for a particular model of consciousness has been chosen for some class of physical systems P, wherein $p \in P$ represents the dynamics or configurations of a particular physical system. $\mathcal{O}$ then denotes all datasets that can arise from observations or measurements on P. Measuring the observables of p maps to datasets $o \in \mathcal{O}$, which is denoted by the obs correspondence. E represents the mathematical description of experience given by the theory or model of consciousness under consideration. In the simplest case, this is just a set whose elements indicate whether a stimulus has been perceived consciously or not, but far more complicated structures can arise (e.g., in IIT). The correspondence $pred$ describes the process of prediction as a map from $\mathcal{O}$ to E.

experimental participant reporting on the switching of some perceptually bistable image (Blake, Brascamp, & Heeger, 2014) or on reports about seen vs. unseen images in masking paradigms (Alais, Cass, O'Shea, & Blake, 2010).

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

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

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

Defining inf as a function means that we assume that for every experimental dataset o, one single experience in E is inferred during the experiment. Here we use a function instead of a correspondence for technical and formal ease, which does not affect our results: If two correspondences to the same space are given, one of them can be turned into a function.[2] The inf function is flexible enough to encompass both direct

[2] If inf is a correspondence, one defines a new space E' by $E' := \{inf(o) \,|\, o \in \mathcal{O}\}$. Every individual

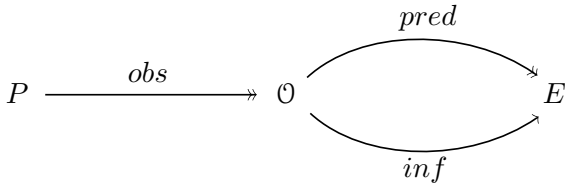


Figure 5.2.2.: Two maps are necessary for a full experimental setup, one that describes a theory's predictions about experience ($pred$), another that describes the experimenter's inference about it (inf). Both map from a dataset $o \in O$ collected in an experimental trail to some subset of experiences described by the model, E.

report, no-report, input/output analysis, and also assumed-report cases. It is a mapping that describes the process of inferring the conscious experience of a system from data recorded in the experiments. Both inf and $pred$ are depicted in Figure 5.2.2.

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

5.2.4. Falsification

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

Definition 5.2.1. *There is a* falsification *at* $o \in \mathfrak{O}$ *if we have*

$$inf(o) \notin pred(o)\,. \tag{5.2}$$

This definition can be spelled out in terms of individual components of E. To this end, for any given dataset $o \in \mathfrak{O}$, let $e_r := inf(o)$ denote the experience that is being inferred, and let $e_p \in obs(o)$ be one of the experiences that is predicted based off of some dataset. Then (5.2) simply states that we have $e_p \neq e_r$ for all possible predictions $e_p \in obs(o)$. None of the predicted states of experience is equal to the inferred experience.

What does Equation (5.2) mean? There are two cases which are possible. Either, the prediction based on the theory of consciousness is correct and the inferred experience is wrong. Or the prediction is wrong, so that in this case the model would be falsified. In short: Either the prediction process or the inference process is wrong.

element of this space describes exactly what can be inferred from one dataset $o \in \mathfrak{O}$, so that $inf' : \mathfrak{O} \to E'$ is a function. The correspondence obs is then redefined, for every $e' \in E'$, by the requirement that $e' \in obs'(o)$ iff $e \in obs(o)$ for some $e \in e'$.

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

In order to understand in more detail what is going on if (5.2) holds, we have to look into a single dataset $o \in \mathcal{O}$. This will be of use later.

Generally, inf and obs will make use of different part of the data obtained in an experimental trial. E.g., in the context of IIT or GNW, data about the internal structure and state of the brain will be used for the prediction. This data can be obtained from an fMRI scan or EEG measurement. The state of consciousness on the other hand can be inferred from verbal reports. Pictorially, we may represent this as in Figure 5.2.3. We use the following notation:

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

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

Note that in both cases, the subscript can be read similarly as the notation for restricting a set. We remark that a different kind of prediction could be considered as well, where one makes use of the inverse of $pred$. In Appendix 5.B, we prove that this is in fact equivalent to the case considered here, so that Definition 5.2.1 indeed covers the most general situation.

5.2.5. Summary

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

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

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

$\mathcal{O}$ is the class of all possible datasets that can result from observations or measurements of the systems in the class P. Any single experimental trail results in

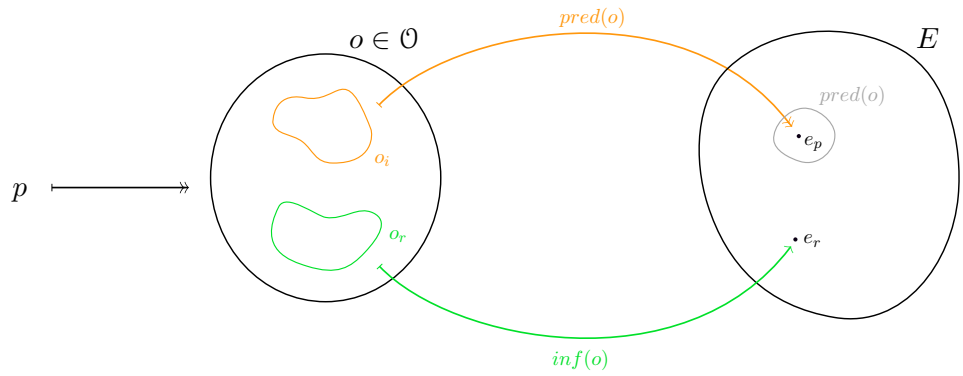


Figure 5.2.3.: This figure represents the same setup as Figure 5.2.2. The left circle depicts one single dataset o. o_i (orange) is the part of the dataset used for prediction. o_r (green) is the part of the dataset used for inferring the state of experience. Usually the green area comprises verbal reports or button presses, whereas the orange area comprises the data obtained from brain scans. The right circle depicts the experience space E of a theory under consideration. e_p denotes a predicted experience while e_r denotes the inferred experience. Therefore, in total, to represent some specific experimental trial we use $p \in P$, $o \in \mathcal{O}$, $e_r \in E$ and $e_p \in E$, where $e_p \in pred(o)$.

a single dataset $o \in O$, whose data is used for making predictions based on the theory of consciousness and for inference purposes.

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

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

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

5.3. The Substitution Argument

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

Here we show that, in general, the contemporary notion of falsification in the science of consciousness exhibits this fundamental flaw for almost all contemporary theories, rather than being a problem for a particular theory. This flaw is based on the independence between the data used for inferences about consciousness (like reports) and the data used to make predictions about consciousness. We discuss various responses to this flaw in Section 5.5.

We begin by defining what a substitution is in Section 5.3.1, show that it implies falsifications in Section 5.3.2, and analyze the particularly problematic case of universal substitutions in Section 5.3.3. In Section 5.3.4, we prove that universal substitutions exist if prediction and inference data are independent and give some examples of already-known cases.

5.3.1. Substitutions

In order to define formally what a substitution is, we work with the inference content o_r of a dataset o as introduced in Section 5.2.4. We first denote the class of all physical configurations which could have produced the inference content o_r upon measurement by P_{o_r}. Using the correspondence obs which describes the relation between physical systems and measurement results, this can be defined as

$$P_{o_r} := \{\, p \in P \,|\, o_r \in obs(p) \,\}\,, \tag{5.3}$$

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

Any map of the form $S : P_{o_r} \to P_{o_r}$ takes a system configuration p which can produce inference content o_r to another system's configuration $S(p)$ which can produce the same inference content. This allows us to define what a substitution is formally. In what follows, the $\circ$ indicates the composition of the correspondences obs and $pred$ to give a correspondence from P to E, which could also be denoted as $pred(obs(p))$,[3] and $\cap$ denotes the intersection of sets.

Definition 5.3.1. *There is a* o_r-substitution *if there is a transformation* $S : P_{o_r} \to P_{o_r}$ *such that at least for one* $p \in P_{o_r}$

$$pred \circ obs(p) \,\cap\, pred \circ obs(S(p)) = \emptyset\,. \tag{5.4}$$

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

[3] I.e., $pred \circ obs(p) = \{e \in E \,|\, e \in pred(o) \text{ for some } o \in obs(p)\}$, it is the image under $pred$ of the set $obs(o)$.

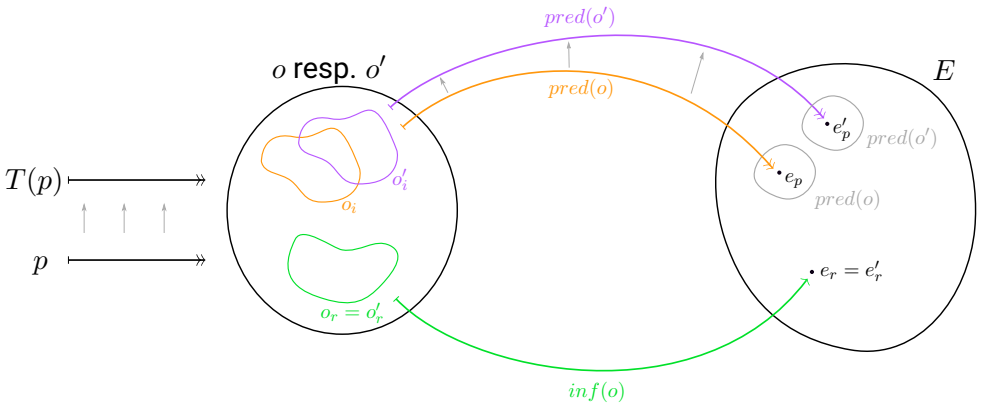


Figure 5.3.1.: This picture illustrates substitutions. Assume that some dataset o with inference content o_r is given. A substitution is a transformation S of physical systems which leaves the inference content o_r invariant but which changes the result of the prediction process. Thus whereas p and $S(p)$ have the same inference content o_r, the prediction content of experimental datasets is different; different in fact to such an extent that the predictions of consciousness based on these datasets are incompatible (illustrated by the non-overlapping gray circles on the right). Here we have used that by definition of P_{o_r}, every $\tilde{p} \in P_{o_r}$ yields at least one dataset o' with the same inference content as o and have identified o and o' in the drawing.

for at least one system configuration p that gives this inference data. In other words, the transformation S is such that for at least one p, the *predictions change completely*, while the inference content o_r is preserved.

A pictorial definition of substitutions is given in Figure 5.3.1. We remark that if $pred$ and obs were functions, so that $pred \circ obs(p)$ only contained one element, Equation (5.4) would be equivalent to $pred(obs(p)) \neq pred(obs(S(p)))$.

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

Definition 5.3.2. *There is a* universal substitution *if there is an o_r-substitution $S_{o_r} : P_{o_r} \to P_{o_r}$ for every o_r.*

We recall that according to the notation introduced in Section 5.2.4, the inference content of any dataset $o \in \mathcal{O}$ is denoted by o_r (adding the subscript r). Thus the requirement is that there is an o_r-substitution $S_{o_r} : P_{o_r} \to P_{o_r}$ for every inference data that can pertain in the experiment under consideration (for every inference data that is listed in $\mathcal{O}$). The subscript o_r of S_{o_r} indicates that the transformation S in Definition 5.3.1 can be chosen differently for different o_r. Definition 5.3.2 does not require there to be one single transformation that works for all o_r.

5.3.2. Substitutions imply falsifications

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

Lemma 5.3.3. *If there is a o_r-substitution, there is a falsification at some $o \in \mathcal{O}$.*

Proof. Let p be the physical system in Definition 5.3.1 and define $p' = S(p)$. Let $o \in obs(p)$ be a dataset of p which has inference content o_r and let o' be a dataset of p' which has the same inference content o_r, guaranteed to exist by the definition of P_{o_r} in (5.3). Equation (5.4) implies that

$$pred(o) \cap pred(o') = \emptyset \,. \tag{5.5}$$

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

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

5.3.3. Universal substitutions imply complete falsification

In Section 5.2.4, we have defined falsifications for individual datasets $o \in \mathcal{O}$. Using the 'insight view' of single datasets, we can refine this definition somewhat by relating it to the inference content only.

Definition 5.3.4. *There is an o_r-falsification if there is a falsification for some $o \in \mathcal{O}$ which has inference content o_r.*

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

Lemma 5.3.5. *If there is an o_r-substitution, there is an o_r-falsification.*

Proof. This lemma follows directly from the proof of Lemma 5.3.3 because the datasets o and o' used in that proof both have inference content o_r. □

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

Proposition 5.3.6. *If there is a universal substitution, there is an o_r-falsification* for all possible *inference contents o_r.*

Proof. By definition of universal substitution, there is an o_r-substitution for every o_r. Thus the claim follows directly from Lemma 5.3.5. □

In combination with Definition 5.3.4, this proposition states that for every possible report (or any other type of inference procedure, cf. our use of terminology in Section 5.2.4), there is a dataset o which contains the report's data and for which we have

$$inf(o_r) \notin pred(o)\,, \tag{5.6}$$

where we have slightly abused notation in writing $inf(o_r)$ instead of $inf(o)$ for clarity. This implies that one of two cases needs to pertain: Either at least one of the inferred experiences $inf(o_r)$ is correct, in which case the corresponding prediction is wrong and the theory needs to be considered falsified. The only other option is that for *all* inference contents o_r, the prediction $pred(o)$ is correct, which qua (5.6) implies that no single inference $inf(o_r)$ points at the correct experience, so that the inference procedure is completely wrong. This shows that Proposition 5.3.6 can equivalently be stated as follows.

Proposition 5.3.7. *If there is a universal substitution, either every single inference operation is wrong or the theory under consideration is already falsified.*

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

5.3.4. When does a universal substitution exist?

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

5.3.4.1. Theories need to be minimally informative

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

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

$$\bar{o} = \{\, o' \,|\, \exists\, p \text{ such that } o \in obs(p) \text{ and } o' \in obs(p) \,\}\,. \tag{5.7}$$

Using this, we can state our *minimal information assumption* in a way that is compatible with the general setup displayed in Figure 5.2.2:

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

$$pred(\bar{o}) \cap pred(\bar{o}') = \emptyset \,. \tag{5.8}$$

5.3.4.2. Inference and prediction data are independent

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

Definition 5.3.8. Inference and prediction data are independent *if for any* o_i, o_i' *and* o_r, *there is a variation*

$$\nu : P \to P \tag{5.9}$$

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

Here, we use the same shorthand as in (5.3). For example, the requirement $o_i \in obs(p)$ is a shorthand for there being an $o \in obs(p)$ which has prediction content o_i. The variation ν in this definition is a variation in P, which describes physical systems which could, in principle, have been realized in an experiment (cf. Section 5.2.5). We note that a weaker version of this definition can be given which still implies our results below, cf. Appendix 5.A. Note that if inference and prediction data are not independent, e.g. because they have a common cause, problems of tautologies loom large, cf. Section 5.5. Throughout the text we often refer to Definition 5.3.8 simply as "independence".

5.3.4.3. Universal substitutions exist

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

Proposition 5.3.9. *If inference and prediction data are independent, universal substitutions exist.*

Proof. To show that a universal substitution exists, we need to show that for every $o \in \mathcal{O}$, an o_r-substitution exists (Definition 5.3.1). Thus assume that an arbitrary $o \in \mathcal{O}$ is given. The minimal information assumption guarantees that there is an o' such that Equation (5.8) holds. As before, we denote the prediction content of o and o' by o_i and o_i', respectively, and the inference content of o by o_r.

Since inference and prediction data are independent, there exists a $p \in P$ as well as a $\nu : P \to P$ such that $o_i \in obs(p)$, $o_i' \in obs(\nu(p))$, $o_r \in obs(p)$ and $o_r \in obs(\nu(p))$. By Definition (5.7), the first two of these four conditions imply that $obs(p) \subset \bar{o}$ and $obs(\nu(p)) \subset \bar{o}'$. Thus Equation (5.8) applies and allows us to conclude that

$$pred(obs(p)) \cap pred(obs(\nu(p)) = \emptyset \,.$$

Via Equation (5.3), the latter two of the four conditions imply that $p \in P_{o_r}$ and $\nu(p) \in P_{o_r}$. Thus we may restrict ν to P_{o_r} to obtain a map

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

which in light of the first part of this proof exhibits at least one $p \in P_{o_r}$ which satisfies (5.4). Thus we have shown that an o_r-substitution exists. Since o was arbitrary, it follows that a universal substitution exists. □

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

Combining this result with Proposition 5.3.7, we finally can state our main theorem.

Theorem 5.3.10. *If inference and prediction data are independent, either every single inference operation is wrong or the theory under consideration is already falsified.*

Proof. The theorem follows by combining Proposition 5.3.9 and Proposition 5.3.7. □

In the next section, we give several examples of universal substitutions, before discussing various possible responses to our result in Section 5.5.

5.3.4.4. Examples of data independence

Our main theorem shows that testing a theory of consciousness will necessarily lead to its falsification if inference and prediction data are independent (Definition 5.3.8), and if at least one single inference can be trusted (Theorem 5.3.10). In this section, we give several examples that illustrate the independence of inference and prediction data. We take report to mean output, behavior, or verbal report itself and assume that prediction data derives from internal measurements.

Artificial neural networks. ANNs, particularly those trained using deep learning, have grown increasingly powerful and capable of human-like performance (LeCun et al., 2015; Bojarski et al., 2016). For any ANN, report (output) is a function of node states. Crucially, this function is non-injective, i.e. some nodes are not part of the output. E.g., in deep learning, the report is typically taken to consist of the last layer of the ANN, while the hidden layers are not taken to be part of the output. Correspondingly, for any given inference data, one can construct a ANN with arbitrary prediction data by adding nodes, changing connections and changing those nodes which are not part of the output. Put differently, one can always substitute a given ANN with another with different internal observables but identical or near-identical reports. From a mathematical perspective it is well-known that both feed-forward ANNs and recurrent ANNs can approximate any given function (Hornik et al., 1989; Schäfer & Zimmermann, 2006). Since reports are just some function, it follows that there are viable universal substitutions.

A special case thereof is the unfolding transformation considered in (Doerig et al., 2019) in the context of IIT. The arguments in this paper constitute a proof of the fact that for ANNs, inference and prediction data are independent (Definition 5.3.8). Crucially, our main theorem shows that this has implications for all minimally informative theories of consciousness. A similar result (using a different characterization of theories of consciousness than minimally informative) has been shown in (Kleiner, 2020a).

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

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

The above are all real and viable classes of systems that are used everyday in science and engineering which all provide different viable universal substitutions if inferences are based on reports or outputs. They show that in normal experimental setups such as the ones commonly used in neuroscientific research into consciousness (Frith, Perry, & Lumer, 1999), inference and prediction data are indeed independent, and dependency is not investigated nor properly considered. It is always possible to substitute the physical system under consideration with another that has different internal observables, and therefore different predictions, but similar or identical reports. Indeed, recent research in using the work introduced in this work shows that even different spatiotemporal models of a system can be substituted for one another, leading to falsification (Hanson & Walker, 2021). We have not considered possible but less reasonable examples of universal substitutions, like astronomically-large look-up ledgers of reports.

As an example of our Main Theorem 5.3.10, we consider the case of IIT. Since the theory is normally applied in Boolean networks, logic gates, or artificial neural networks, one usually takes report to mean "output." In this case, it has already been proven that systems with different internal structures and hence different predicted experiences, can have identical input/output (and therefore identical reports or inferences about report) (Albantakis & Tononi, 2019). To take another case: within IIT it has already been

acknowledged that a Turing machine may have a wildly different predicted contents of consciousness for the same behavior or reports (C. Koch, 2019). Therefore, data independence during testing has already been shown to apply to IIT under its normal assumptions.

5.4. Inference and Prediction Data are Strictly Dependent

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

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

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

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

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

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

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

Definition 5.4.1. *A theory of consciousness which does not have a falsification over $\overline{\mathcal{O}}$ is* empirically unfalsifiable*.*

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

Definition 5.4.2. *Inference and prediction data are* strictly dependent *if there is a function f such that for any $o \in \overline{\mathcal{O}}$, we have $o_i = f(o_r)$.*

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

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

$$o_i \subseteq o_r \, . \tag{5.10}$$

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

In all these cases, consciousness is generated by – and hence needs to be predicted via – what is accessible to report or output. In terms of Block's distinction between phenomenal consciousness and access consciousness (Block, 1996), Equation (5.10) holds true whenever a theory of consciousness is under investigation where access consciousness determines phenomenal consciousness.

Our second main theorem is the following.

Theorem 5.4.3. *If a theory of consciousness implies that inference and prediction data are strictly dependent, then it is either already falsified or empirically unfalsifiable.*

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

$$\mathcal{O}_{\mathrm{exp}} := \{ \, (o_i, o_r) \mid o \in \mathcal{O} \, \} \, , \tag{5.11}$$

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

$$\mathcal{O}_{\text{all}} := \{\, (o_i, o'_r) \mid o \in \mathcal{O},\ o' \in \mathcal{O} \} \tag{5.12}$$

$$= \{\, (o_i, o'_r) \mid o_i \in \mathcal{O}_i,\ o'_r \in \mathcal{O}_r \,\}\,. \tag{5.13}$$

Crucially, here, o_i and o'_r do not have to be part of the same dataset o. Combined with Definition 5.2.1, we conclude that there is a falsification over $\overline{\mathcal{O}}$ if for some $(o_i, o'_r) \in \mathcal{O}_{\text{all}}$, we have $inf(o) \notin pred(o')$, and there is a falsification over $\mathcal{O}$ if for some $(o_i, o_r) \in \mathcal{O}_{\text{exp}}$, we have $inf(o) \notin pred(o)$.

Next we show that if inference and prediction data are strictly dependent, then $\mathcal{O}_{\text{all}} = \mathcal{O}_{\text{exp}}$ holds. We start with the set $\mathcal{O}_{\text{all}}$ as defined in (5.12). Expanding this definition in words, it reads

$$\mathcal{O}_{\text{all}} = \{\, (d_i, d_r) \mid \exists\, o \in \mathcal{O} \text{ such that } d_r = o_r \text{ and } \exists\, \tilde{o} \in \mathcal{O} \text{ such that } d_i = \tilde{o}_i \,\}\,, \tag{5.14}$$

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

Assume that the first condition in this expression, $d_r = o_r$ holds for some $o \in \mathcal{O}$. Since inference and prediction data are strictly dependent, we have $d_i = f(d_r)$. Furthermore, for the same reason, the prediction content o_i of the dataset o satisfies $o_i = f(o_r)$. Applying the function f to both sides of the first condition gives $f(d_r) = f(o_r)$, which thus in turn implies $o_i = d_i$. This means that the o that satisfies the first condition in (5.14) automatically also satisfies the second condition. Therefore, due to inference and prediction data being strictly dependent, (5.14) is equivalent to

$$\mathcal{O}_{\text{all}} = \{\, (d_i, d_r) \mid \exists\, o \in \mathcal{O} \text{ such that } d_r = o_r \text{ and } d_i = o_i \,\}\,. \tag{5.15}$$

This, however, is exactly $\mathcal{O}_{\text{exp}}$ as defined in (5.11). Thus we conclude that if inference and prediction data are strictly dependent, $\mathcal{O}_{\text{all}} = \mathcal{O}_{\text{exp}}$ necessarily holds.

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

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

of empirical evidence required in any scientific investigation. Thus such theories are empirically unfalsifiable.

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

5.5. Objections

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

5.5.1. Restricting inferences to humans only

The examples given in Section 5.3.4.4 show that data independence holds during the usual testing setups. This is because prima facie it seems reasonable to base inferences either on report capability or intelligent behavior in a manner agnostic of the actual physical makeup of the system. Yet this entails independence, so in these cases our conclusions apply.

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

Overall there are strong reasons to reject this restriction of inferences. One significant issue is that this objection is equivalent to saying that reports or behavior in non-humans carry *no* information about consciousness, an incredibly strong claim. If non-humans contradicted a theory (like a complex organism acting in pain while a theory predicted a lack of pain) the theory would be presumed to be correct above any behavior or report, meaning that abductive application of the theory ignores the fact that this sort of abductive reasoning should actually falsify the theory. Indeed, this is highly problematic for consciousness research which often uses non-human animal models (Boly et al.,

2013). For instance, cephalopods are among the most intelligent animals yet are quite distant on the tree of life from humans and have a distinct neuroanatomy, and still are used for consciousness research (Mather, 2008). Even in artificial intelligence research, there is increasing evidence that deep neural networks produced brain-like structures such as grid cells, shape tuning, and visual illusions, and many others (Richards et al., 2019). Given these similarities, it becomes questionable why the strength of inferences should be based on homology instead of capability of report or intelligence.

What is more, restricting inferences to humans alone is unlikely to be sufficient to avoid our results. Depending on the theory under consideration, data independence might exist just in human brains alone. That is, it is probable that there are transformations (as in Equation (5.9)) available within the brain wherein o_r is fixed but o_i varies. This is particularly true once one allows for interventions on the human brain by experimenters, such as perturbations like transcranial magnetic stimulation, which is already used in consciousness research (Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010; Napolitani et al., 2014).

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

5.5.2. Reductio ad absurdum

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

The assumption in question is that $\mathcal{O}$, the class of all datasets that can result from observations or measurements of a system, is determined by the physical configurations in P alone. I.e., every single dataset o, including both its prediction content o_i and its inference content o_r, is determined by p, and *not* by a conscious experience in E. In Figure 5.2.2, this is reflected in the fact that there is an arrow from P to $\mathcal{O}$, but no arrow from E to $\mathcal{O}$.

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

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

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

$$P \longrightarrow \mathcal{O} \longleftarrow T \,,$$

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

5.5.3. Theories could be based on phenomenology

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

In addition to elegance and parsimony, in consciousness science, one could in particular consider a theory's fit with phenomenology, i.e. how well a theory describes the general structure of conscious experience. Examples of theories that are constructed based on a fit with phenomenology are recent versions of IIT (Oizumi et al., 2014) or any view that proposes developing theories based on isomorphisms between the structure of experiences and the structure of physical systems or processes (Tsuchiya et al., 2020).

It might be suggested that phenomenological theories might be immune to aspects of the issues we outline in our results (Negro, 2020). We emphasize that in order to avoid our results, and indeed the need for any experimental testing at all, a theory constructed from phenomenology has to be *uniquely derivable* from conscious experience. However, to date, no such derivation exists, as phenomenology seems to generally underdetermine the postulates of IIT (Bayne, 2018; A. B. Barrett & Mediano, 2019), and be-

cause it is unknown what the scope and nature of non-human experience is. Therefore theories based on phenomenology can only confidently identify systems with human-like conscious experiences and cannot currently do so uniquely. Thus they cannot avoid the need for testing.

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

5.5.4. Rejecting falsifiability

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

It is important to note, however, that Lakatos does not actually break with falsificationism. This is why Lakatos description of science is often called "refined falsificationism" in philosophy of science (Radnitzky, 1991). Thus cases of testing theories' predictions remain relevant in a Lakatosian view in order to distinguish between progressive and degenerating research programs. Therefore our results generally translate into this view of scientific progress. In particular, Theorem 5.3.10 shows that for every single inference procedure that is taken to be valid, there exists a system for which the theory makes a wrong prediction. This implies necessarily that a research program is degenerating. That is, independence implies that there is always an available substitution that can falsify any particular prediction coming from the research program.

5.6. Conclusion

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

Crucially, in contrast to other similar results (Doerig et al., 2019), we do not put the blame on individual theories of consciousness, but rather show that a key assumption that is usually being made is responsible for the problems: an experimenter's inference about consciousness and a theory's predictions are generally implicitly assumed to be independent during testing across contemporary theories. As we formally prove, if this independence holds, substitutions or changes to physical systems are possible that

falsify any given contemporary theory. Whenever there is an experimental test of a theory of consciousness on some physical system which does not lead to a falsification, there necessary exists another physical system which, if it had been tested, would have produced a falsification of that theory. We emphasize that this problem does not only affect one particular type of theory, for example those based on causal interactions like IIT; theorems apply to all contemporary neuroscientific theories of consciousness if independence holds.

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

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

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

A priori, in the space of possible relationships between inference and prediction data, there seems to be room for relationships that are neither independent (Section 5.3) nor strictly dependent (Section 5.4). We define this relationships of this kind as cases of *lenient dependency*. No current theory or testing paradigm that we know of satisfies this definition. Yet cases of lenient dependency cannot be excluded to exist. Such cases would technically not be beholden to either Theorem 5.3.10 or Theorem 5.4.3.

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

Physics is not causally closed. Another way to avoid our conclusion is to only consider theories of consciousness which do not describe the physical as causally closed (Kim, 1998). That is, the presence or absence of a particular experience itself would have to make a difference to the configuration, dynamics, or states of physical systems above

and beyond what would be predicted with just information about the physical system itself. If a theory of consciousness does not describe the physical as closed, a whole other range of predictions are possible: predictions which concern the physical domain itself, e.g., changes in the dynamics of the system which depend on the dynamics of conscious experience. These predictions are not considered in our setup and may serve to test a theory of consciousness without the problems we have explored here.

5.7. Acknowledgments

We would like to thank David Chalmers, Ned Block, and the participants of the NYU philosophy of mind discussion group for valuable comments and discussion. Thanks also to Ryota Kanai, Jake Hanson, Stephan Sellmaier, Timo Freiesleben, Mark Wulff Carstensen and Sofiia Rappe for feedback on early versions of the manuscript.

Author contributions: J.K.and E.H. conceived the project and wrote the article.
Competing interests: The authors declare no competing interests.

Appendix

5.A. Weak Independence

In this section, we show how Definition 5.3.8 can be substantially relaxed while still ensuring our results to hold. To this end, we need to introduce another bit of formalism: We assume that predictions can be compared to establish how different they are. This is the case, e.g., in IIT where predictions map to the space of maximally irreducible conceptual structures (MICS), sometimes also called the space of Q-shapes, which carries a distance function analogous to a metric (Kleiner & Tull, 2021). We assume that for any given prediction, one can determine which of all those predictions that don't overlap with the given one is *most similar* to the latter, or equivalently which is *least different*. We calls this a *minimally differing prediction* and use it to induce a notion of *minimally differing data sets* below. Uniqueness is not required.

Let an arbitrary data set $o \in \mathcal{O}$ be given. The minimal information assumption from Section 5.3.4.1 ensures that there is at least one data set o' such that Equation (5.8) holds. For what follows, let $o^{\perp}$ denote the set of all data sets which satisfy Equation (5.8) with respect to o,

$$o^{\perp} := \{\, o' \in \mathcal{O} \mid pred(\bar{o}) \cap pred(\bar{o}') = \emptyset \,\}\,. \tag{5.16}$$

Thus $o^{\perp}$ contains all data sets whose prediction completely differs from the prediction of o.

Definition 5.A.1. *We denote by* $\min(o)$ *those data sets in* $o^{\perp}$ *whose prediction is least different from the prediction of* o*.*

In many cases $\min(o)$ will only contain one data set, but here we treat the general case where this is not so. We emphasize that the minimal information assumption guarantees that $\min(o)$ exists. We can now specify a much weaker version of Definition 5.3.8.

Definition 5.A.2. Inference and prediction data are independent *if for any* $o \in \mathcal{O}$ *and* $o' \in \min(o)$*, there is a variation*

$$\nu : P \to P \tag{5.17}$$

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

The difference between Definition 5.A.2 and Definition 5.3.8 is that for a given $o \in \mathcal{O}$, the latter requires the transformation ν to exist for any $o' \in \mathcal{O}$, wheres the former only requires it to exist for minimally different data sets $o' \in \min(o)$. The corresponding proposition is the following.

Proposition 5.A.3. *If inference and prediction data are weakly independent, universal substitutions exist.*

Proof. To show that a universal substitution exists, we need to show that for every $o \in \mathcal{O}$, an o_r-substitution exists (Definition 5.3.1). Thus assume that an arbitrary $o \in \mathcal{O}$ is given and pick an $o' \in \min(o)$. As before, we denote the prediction content of o and o' by o_i and o'_i, respectively, and the inference content of o by o_r.

Since inference and prediction data are weakly independent, there exists a $p \in P$ as well as a $\nu : P \to P$ such that $o_i \in obs(p)$, $o'_i \in obs(\nu(p))$, $o_r \in obs(p)$ and $o_r \in obs(\nu(p))$. By Definition (5.7), the first two of these four conditions imply that $obs(p) \subset \bar{o}$ and $obs(\nu(p)) \subset \bar{o}'$. Since o' is in particular an element of $o^{\perp}$, Equation (5.8) applies and allows us to conclude that

$$pred(obs(p)) \cap pred(obs(\nu(p)) = \emptyset \, .$$

Via Equation (5.3), the latter two of the four conditions imply that $p \in P_{o_r}$ and $\nu(p) \in P_{o_r}$. Thus we may restrict ν to P_{o_r} to obtain a map

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

which in light of the first part of this proof exhibits at least one $p \in P_{o_r}$ which satisfies (5.4). Thus we have shown that an o_r-substitution exists. Since o was arbitrary, it follows that a universal substitution exists. □

The following theorem shows that Definition 5.A.2 is sufficient to establish the claim of Theorem 5.3.10.

Theorem 5.A.4. *If inference and prediction data are weakly independent, either every single inference operation is wrong or the theory under consideration is already falsified.*

Proof. The theorem follows by combining Proposition 5.A.3 and Proposition 5.3.7. □

5.B. Inverse Predictions

When defining falsification, we have considered predictions that take as input data about the physical configuration of a system and yield as output a state of consciousness. An alternative would be to consider the inverse procedure: a prediction which takes as input a reported stated of consciousness and yields as output some constraint on the physical configuration of the system that is having the conscious experience. In this section, we discuss the second case in detail.

As before, we assume that some data set o has been measured in an experimental trail, which contains both the inference data o_r (which includes report and behavioural indicators of consciousness used in the experiment under consideration) as well as some data o_i that provides information about the physical configuration of the system

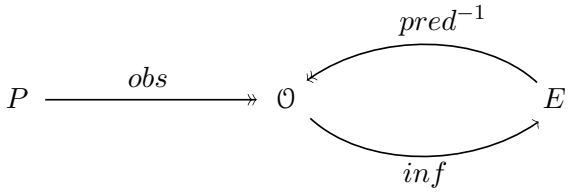


Figure 5.B.1.: The case of an inverse prediction. Rather than comparing the inferred and predicted state of consciousness, one predicts the physical configuration of a system based on the system's report and compares this with measurement results.

under investigation. For simplicity, we will also call this *prediction data* here. Also as before, we take into account that the state of consciousness of the system has to be inferred from o_r, and again denote this inference procedure by inf.

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

$$pred^{-1}(e) = \{\, o \in \mathcal{O} \,|\, e \in pred(o) \} \,. \tag{5.18}$$

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

$$pred^{-1}\big(inf(o)\big) \,, \tag{5.19}$$

depicted in Figure 5.B.1, and a falsification occurs in case the the observed o has a prediction content o_i which is not in this set. Referring to the previous definition of falsification as *type-1* (Definition 5.2.1), we define this new form of falsification as *type-2*.

Definition 5.B.1. *There is a* type-2 falsification *at $o \in \mathcal{O}$ if we have*

$$o \notin pred^{-1}\big(inf(o)\big) \,. \tag{5.20}$$

In terms of the notion introduced in Section 5.2.5, Equation (5.20) could equivalently be written as $o_i \notin pred^{-1}\big(inf(o_r)\big)_i$. The following lemma shows that there is a type-2 falsification if and only if there is a type-1 falsification. Hence all of our previous results apply as well to type-2 falsifications.

Lemma 5.B.2. *There is a type-2 falsification at o if and only if there is a type-1 falsification at o.*

Proof. Equation (5.18) implies that $o \notin pred^{-1}(e)$ if and only if $e \notin pred(o)$. Applied to $e = inf(o)$, this implies:

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

The former is the definition of a type-2 falsification. The latter is Equation (5.2) in the definition of a type-1 falsification. Hence the claim follows. □

6. The Closure of the Physical, Consciousness and Scientific Practice

Johannes Kleiner, Stephan Hartmann

6.1. Introduction

The closure of the physical is a central assumption in the philosophy of mind and in the scientific study of consciousness (Kim, 1996; Papineau, 2009). It underlies both functionalist and identity theories of consciousness and is a central component of many, if not all, neuroscientific models of consciousness. However, we will show below that the closure of the physical is untenable in a scientific context because it implies that no experiment can actually distinguish between two theories of consciousness that obey this assumption. It is therefore incompatible with scientific practice and hence *unscientific*.

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

It has already been pointed out that the closure of the physical is a problematic assumption in a scientific context. (Pauen, 2000) and (Pauen, 2006), for example, make this point with respect to property dualism and qualia epiphenomenalism. Our proof presented below covers the general case. It shows independently of any other metaphysical premises that one of the central assumptions in the empirical study of consciousness is flawed. This calls into question the theoretical basis of a large number

of experiments conducted to date and shows that the hope of basing a physical functionalist or identity-based understanding of consciousness on empirical observations is null and void.

The remainder of this paper is organized as follows. Section 6.2 elaborates which theories of consciousness our argument addresses and defines an epistemic version of the closure of the physical. Section 6.3 identifies a necessary condition for theories of consciousness to be distinguished by empirical data. Sections 6.4 and 6.5 discuss the role of empirical data in the scientific study of consciousness and why they supervene on physical events. Section 6.6 is devoted to the proof of our main claim, and Section 6.7 shows that the causal closure of the physical, as usually defined ontologically, implies our definition, which ensures that our result holds in full generality. Finally, Section 6.8 contains some concluding remarks.

6.2. Theories of Consciousness

We use the term *theories of consciousness* to refer to the theories that are tested, compared, or derived in experiments in the scientific study of consciousness, regardless of what metaphysical status of consciousness they presuppose. This includes, for example, Integrated Information Theory (Oizumi et al., 2014), Global Neuronal Workspace Theory (Mashour, Roelfsema, Changeux, & Dehaene, 2020) or Higher Order Thought Theory (Brown, Lau, & LeDoux, 2019), and in general all scientific theories which adhere to functionalism, identity theory or epiphenomenalism. This also includes illusionist or eliminativist theories that are subject to experimental testing, even though they do not grant consciousness an independent ontological status, but merely aim to explain why someone has the illusion of being conscious (Sprevak & Irvine, 2020).

Our results rely on two general facts about theories of consciousness. The first is that theories of consciousness relate to physical events, where *physical events* are the kinds of events that are the subject of natural sciences such as biology, chemistry, neuroscience, and physics. Some theories modify the description of physical events provided by natural science, for example, by postulating changes in the temporal evolution of physical states, as recently in (Chalmers & McQueen, 2022), others simply adopt whatever natural science says about physical events without any modification.

The causal closure of the physical is the assumption that for every physical effect, there is a sufficient physical cause. Its key epistemic repercussion (cf. Section 6.7) is that theories of consciousness must not amend whatever it is that the physical sciences say or imply about physical events. We call this epistemic assumption *closure of the physical*: A theory of consciousness obeys the *closure of the physical* if and only if it does not posit any changes to the physical events explained, predicted or otherwise determined by natural science.

This premise can be expressed concisely in formal terms. To this end, we introduce two sets[1] of event-descriptions. First, for any theory of consciousness T, we denote by

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

$\mathcal{P}_T$ the physical events which T is committed to, for example the firing of some neurons or the instantiation of some functional property. Every element in $\mathcal{P}_T$ is a description of an event that occurs, according to T, in the actual world. The description specifies the event and may include properties or relational information about the event. What exactly a description contains and in which language it is formulated is not of importance here.

Second, we denote by $\mathcal{P}_P$ the physical events which natural science explains, predicts or determines. Whatever it is that natural science says or implies about the physical events in the actual world is part of the class $\mathcal{P}_P$. Each element is in turn a description of an event, including its properties and relations, and we allow that the description is either deterministic or indeterministic.[2]

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

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

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

$$\iota : \mathcal{P}_T \longrightarrow \mathcal{P}_P \,, \tag{6.1}$$

which specifies for every physical event and description that the theory of consciousness is committed to the corresponding event and description explained, predicted, or determined by natural science. The existence of this function is the concise meaning of the closure of the physical introduced above: A theory of consciousness T obeys the *closure of the physical* if and only if there exists a function ι as in (6.1). We will show in Section 6.7 that the usual reading of the causal closure of the physical implies just that.[3]

[2] In terms of a fundamental physical theory, $\mathcal{P}_P$ may be thought of as comprising all events which are part of those dynamically possible trajectories that occur in the actual world.

[3] The closure of the physical so conceived could also be defined in terms of variables and other concepts used in scientific theories, such that a theory of consciousness obeys the closure of the physical if and only if it makes no change to the concepts that natural scientific theories employ to predict and explain physical events, or which otherwise determine physical events. While this formulation would capture the more familiar assumption that "physical laws already form a closed system" (Chalmers, 1996, p. 127), it introduces another level of abstraction (concepts used in scientific theories) that is avoided when formulated in terms of events.

6.3. Experiments

In the scientific study of consciousness, experiments are conducted to falsify, confirm, or distinguish between competing theories of consciousness. The most important component of any experiment is measurement, i.e., laboratory operations that produce a set of data which constitutes the result of the measurement.

The second general fact on which our argument is based is that scientific theories of consciousness have something to say about possible measurement results. We assume that any theory allows one to derive, for some experiments and under appropriate auxiliary assumptions, a class of data sets which, according to the theory, may occur as the result of the experiment. This requirement singles out *scientific* theories as those to which our argument applies.[4]

We use the symbol M to represent an experiment, and furthermore introduce the symbol $\mathcal{O}_M$ to denote all data sets which could result from this experiment according to some assumption or theory. So $\mathcal{O}_M$ denotes the possible measurement results of M in some context. If an experiment M only made measurements on one system and everything were deterministic, then there would only be one data set in $\mathcal{O}_M$. But experiments usually consider many systems and things are not deterministic, which is why we have a whole class of data sets that can occur in M.[5]

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

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

[4] In particular, if we assume that experiments are required to distinguish between competing theories of consciousness, we assume that consciousness cannot be deduced from the physical or, if it can, that experiments are required to figure out how because the deduction fails in practice due to complexity and/or too little knowledge.

[5] For now, M can be thought of as an experiment actually conducted in the actual world to distinguish between theories of consciousness, although logical possibilities will come into play in Section 6.4.

both classes and is thus consistent with one but not with both theories.[6]

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

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

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

Experiments in the scientific study of consciousness usually use two different types of measurements (Chalmers, 2004). First, they make use of what are called *third-person measurements* which employ standard scientific methods. Typical examples are EEG or fMRI recordings. Second, they use what might be called *first-person* or *consciousness-inferring* measurements. This class of measurements has been characterized as using the subject's access to his or her own conscious experience in some way, such as via verbal reports or pressing of a button (Metzinger, 1995). More recently, the term *subjective measures of consciousness* has come to refer to these types of measures (Irvine, 2013), in contrast to *objective measures* and *no-report paradigms* (Tsuchiya et al., 2015), which infer a subject's state of consciousness indirectly, e.g., by evaluating forced choice tasks (Del Cul et al., 2007) or behavioral data such as optokinetic nystagmus and the pupillary reflex (Frässle, Sommer, Jansen, Naber, & Einhäuser, 2014).

What exactly the difference is between measurements in the first and third person is not important for our purposes. The only important thing is that both types of measurements produce results that need to be analyzed, interpreted or transformed. To do this, they must be stored on a data repository. This fact has implications that we analyze below.[7]

[6] Note that empirical indistinguishability is weaker than empirical equivalence, as defined, for example, in (Weatherall, 2019a) and (Weatherall, 2019b). Two theories are empirically indistinguishable if they make exactly the same testable statements about experiments to which they are both applicable. Empirical equivalence also requires that the two theories apply to exactly the same experiments.

[7] We emphasize that this also holds true for "measuring" consciousness by introspection. Because science is an intersubjective endeavor, whatever is accessed by introspection in any experiment that aims to distinguish among competing theories of consciousness has to be stored or transmitted in order to be shared with other scientists. Nothing hinges on how precisely one flashes out what is special about

6.4. Data

We have minimally characterized measurements as laboratory operations that provide a data set that is designated as the result of the experiment. But what does it mean that this data set must be stored on some device? To address this question, let's take a hard disk as an example. A hard disk stores data by magnetizing a thin film of ferromagnetic material that forms the surface of the hard disk platter. The film is made up of many tiny, sequentially aligned magnetic regions, each of which has a magnetization vector that can point in one of two directions. When data is stored on the disk, the head of the drive arm moves over these areas and changes the magnetization vector by applying electric fields. When reading data from the disk, the actuator arm uses weaker electric fields to sense the magnetization vectors of the areas.

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

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

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

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

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

consciousness and its measurement. What matters below is only that measurement results need to be stored or transmitted and that different theories of consciousness may be formulated which are compatible with the same set of physical events. The closure of the physical enforces the latter.

by natural science for the actual world form a subset of $\mathbb{P}$, the subset $\mathcal{P}_P$ we introduced above. The same is true for the physical events $\mathcal{P}_T$ to which a theory of consciousness is committed.

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

$$d_D : P(\mathbb{P}) \longrightarrow P(\mathbb{O}_D) \,, \tag{6.2}$$

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

$$d_D : \mathcal{P} \longmapsto \mathcal{O}_D \,. \tag{6.3}$$

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

In order to use this function in the following, we have to consider two conditions. The first condition arises from the fact that the data stored on a device D corresponding to some physical events is independent of the language used to describe those events. Applied to the embedding ι introduced in (6.1), this means that

$$d_D\big(\iota(\mathcal{P}_T)\big) = d_D\big(\mathcal{P}_T\big) \,. \tag{6.4}$$

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

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

$$d_D\big(\mathcal{P}_1\big) = d_D\big(\mathcal{P}_2\big) \,. \tag{6.5}$$

6.5. Measurement Results

We are now ready to apply this result on data storage to experiments in the scientific study of consciousness. The measurements performed in these experiments tend to

be quite complex. They may employ advanced brain imaging techniques such as EEG, ECoG, or fMRI, and require finely tuned equipment and sophisticated analysis to learn about a subject's state of consciousness.

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

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

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

As before, let us denote by M an arbitrary but fixed experiment in the scientific study of consciousness, and let us denote by D the data store or stores necessarily used in M to store the results of the measurement. We have already introduced the symbol $\mathfrak{O}_M$ to denote the data sets that, under certain assumptions or theories, could be the possible outcomes of the experiment M. Our analysis from the previous section then shows that $\mathfrak{O}_M$ is also determined by the function d_D introduced in (6.2), namely by restricting d_D to the part of the data stored on D that represents the measurement results. If we denote this restriction by d_M and all data sets that could possibly result from M by $\mathbb{O}_M$, we obtain a function

$$\begin{aligned} d_M : P(\mathbb{P}) &\longrightarrow P(\mathbb{O}_M) \\ \mathcal{P} &\longmapsto \mathfrak{O}_M \,, \end{aligned} \tag{6.6}$$

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

The function d_M establishes a connection between what a theory of consciousness T predicts or postulates about physical events in the real world, on the one hand, and the

possible measurement outcomes that can occur according to T, on the other. It selects from the events $\mathcal{P}_T$ that the theory T is committed to those events which determine the data that is stored on D. Making use of the symbol $\mathcal{O}_T$ introduced above to denote the possible measurement results that can occur in M after T, this means that

$$d_M(\mathcal{P}_T) = \mathcal{O}_T \,. \tag{6.7}$$

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

6.6. Why the Closure of the Physical is Unscientific

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

Theorem 6.6.1. *The closure of the physical is unscientific.*

Proof. Let T_1 and T_2 denote two theories of consciousness which obey the closure of the physical. This implies that there exist embeddings $\iota_1 : \mathcal{P}_{T_1} \longrightarrow \mathcal{P}_P$ and $\iota_2 : \mathcal{P}_{T_2} \longrightarrow \mathcal{P}_P$ as in (6.1). Let M denote an experiment to which both T_1 and T_2 are applicable, and D the data storage device(s) used in that experiment. Because of condition (6.4), we have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\mathcal{P}_{T_1})$ and $d_D(\iota_2(\mathcal{P}_{T_2})) = d_D(\mathcal{P}_{T_2})$.

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

Because ι_1 is an embedding, we have $\iota_1(\mathcal{P}_{T_1}) \subset \mathcal{P}_P$. Because D is instantiated in $\iota_1(\mathcal{P}_{T_1})$, Equation (6.5) applies so that we have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\mathcal{P}_P)$. The same applies to ι_2, so that also here, Equation (6.5) implies $d_D(\iota_2(\mathcal{P}_{T_2})) = d_D(\mathcal{P}_P)$. So we in fact have $d_D(\iota_1(\mathcal{P}_{T_1})) = d_D(\iota_2(\mathcal{P}_{T_2}))$, which in light of the above implies $d_D(\mathcal{P}_{T_1}) = d_D(\mathcal{P}_{T_2})$.

We thus find that the data stored on D is exactly the same for both theories. Restriction to d_M introduced in (6.6) furthermore implies that $d_M(\mathcal{P}_{T_1}) = d_M(\mathcal{P}_{T_2})$, and because of (6.7), this implies that $\mathcal{O}_{T_1} = \mathcal{O}_{T_2}$. So the measurement results of M are exactly the same according to both T_1 and T_2. Independently of which predictions one arrives at by making use of auxiliary assumptions, the closure of the physical implies that the data sets which can occur in M cannot differ.

Since M was chosen arbitrarily, this conclusion holds for any experiment M, so T_1 and T_2 are empirically indistinguishable. And because T_1 and T_2 were arbitrarily chosen among the theories obeying the closure of the physical, we can conclude that all theories

obeying the closure of the physical are empirically indistinguishable. It follows that the closure of the physical is an assumption that is unscientific. □

6.7. Causal Closure of the Physical

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

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

According to Jaegwon Kim, for example, the causal closure of the physical implies that "to explain the occurrence of a physical event we never need to go outside of the physical realm" (Kim, 1996, p. 147). And Frank Jackson characterizes the causal closure of the physical as the claim that "the physical sciences, or rather some natural extension of them, can in principle give a complete explanation for each and every bodily movement, or at least can do so up to whatever completeness is compatible with indeterminism in physics" (Jackson, 1996, p. 378).

These statements exemplify that the causal closure of the physical is generally taken to imply that every physical event which is explained at all, is explainable by natural science. But explanation, precisely construed (Strevens, 2006), is only one way in which a theory can address events. Making room for prediction and other possible ways as well, we may take the above to imply that every physical event which is predicted, explained, or determined at all, can be predicted, explained, or determined by natural science.

Applied to a theory of consciousness, this means that any physical event that the theory explains, predicts, or determines can (eventually) be explained, predicted, or determined by natural science. But for this to be true, the theory must not replace, alter, or add to the natural science account of physical events, because otherwise it would be committing itself to physical events that cannot be explained, predicted, or determined by natural science. Thus, the causal closure of the physical implies that a theory of consciousness cannot make changes to the physical events that are explained, predicted, or determined by natural science.

This point can be stated more clearly in formal terms. We have denoted the set of physical events that a theory of consciousness is committed to by $\mathcal{P}_T$. These are the events explained, predicted, or otherwise determined by that theory. And we have denoted the set of physical events explained, predicted, or otherwise determined by natural science (now or in the future) by $\mathcal{P}_P$. Thus, if every physical event that can be explained,

predicted, or determined at all can be explained, predicted, or determined by natural science, then every event that is in $\mathcal{P}_T$ is also in $\mathcal{P}_P$. Taking into account the different languages that can be used in the two cases, this means that for every event description in $\mathcal{P}_T$ there is a corresponding event description of the same event in $\mathcal{P}_P$. This constitutes an injective function that maps $\mathcal{P}_T$ to $\mathcal{P}_P$.

We thus arrive at exactly the same formal requirement as in Equation (6.1). The causal closure of the physical implies that there is an embedding $\iota : \mathcal{P}_T \to \mathcal{P}_P$ that specifies for each physical event and physical description that the theory of consciousness is bound to the corresponding event and description explained, predicted, or determined by natural science.[8] Causal closure of the physical implies closure of the physical, and as a corollary of Theorem 6.6.1 we posit that causal closure of the physical is also unscientific.[9]

We emphasize that nowhere in our argument do we restrict to physical events which are already explained or predicted by natural science. What matters is only which relation a theory of consciousness proposes between the physical events it is committed to and the physical events that natural science posits. Even if a theory presupposes that the physical events it associates with conscious experiences are determined by physical laws, but cannot in practice be explained or predicted based on these laws, as weak emergentist theories would have it, our argument applies. Theories of this sort may be wrong about what they say about physical events, and experiments may help to determine whether this is the case, but insofar as they buy into the very same underlying account of physical events as all other theories, the measurement results necessarily are the same as if any other theory were true. Because of the weak emergence claim, no postulate of such theory can imply any changes in the underlying physical events, and *ipso facto* no changes in measurement results.[10]

[8] More advanced formulations of the causal closure of the physical lead to the same conclusion. Consider for example, the proposal by Barbara Montero and David Papineau in (Montero & Papineau, 2005), that "[e]very physical event is determined, in so far as it is determined at all, by preceding physical conditions and laws". Every physical event that is determined by preceding physical conditions and laws is an element of the class $\mathcal{P}_P$. Every element of $\mathcal{P}_T$ is, according to the broad reading of 'determined' applied (Montero & Papineau, 2005), determined by a theory of consciousness. Hence it follows that every event in $\mathcal{P}_T$ is also in $\mathcal{P}_P$, and taking into account the different languages that may be used to describe the event, that there is an embedding ι as in Equation (6.1).

[9] We note that the commonly understood epistemic reading of the closure of the physical, as expressed in Kim and Jackson's remarks, follows from the causal closure of the physical, as defined in the beginning of this section, only if an appropriate notion of 'physical' is presupposed. This means that the causal closure of the physical must forbid the introduction of new physical entities that have effects that are not explained by the physical sciences.

[10] That is not so for strong emergentist theories, of course. These introduce genuine new causes and effects which are not claimed to be reducable to fundamental physical laws. It is well known that strong emergentist theories are not compatible with physicalism and the causal closure of the physical (O'Connor, 2021a).

6.8. Conclusion

We have shown that the causal closure of the physical goes far beyond what is usually considered. Since all measurement results in the scientific study of consciousness are either physical events (such as keystrokes or sound waves) or at least determined by physical events (such as data stored on hard disks), no two theories obeying the causal closure of the physical can actually be distinguished in experiments. Our result applies to all major neuroscientific theories of consciousness as well as to the leading philosophical paradigms in the field. It applies to any theory of consciousness that fits into the natural science account of physical events without altering it. This includes all functionalist and identity theories of consciousness, such as GNW (Mashour et al., 2020), HOT (Brown et al., 2019), AST (Graziano & Webb, 2015), or predictive processing-based theories (Schlicht & Dolega, 2021), as well as eliminativist or illusionist theories (Frankish, 2016). But it also includes theories such as IIT, whose mathematics takes the form of a function that maps physical states and events to conscious states and events (Kleiner & Tull, 2021).[11]

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

There are two potential conclusions that one can draw from our results. Either, experimenters do not really adhere to the closure of the physical when conducting experiments, but implicitly assume that the theories tested modify what falls solely within the realm of natural science. If this is the case, then our results constitute an imperative to improve the tested theories and make explicit what is implicitly assumed. If, on the other hand, experimenters do not implicitly adhere to the closure of the physical when running experiments, then our results call into question the very conclusions drawn on

[11] Our results do not, however, apply to theories of temperature, life, or similar. They are fully compatible with there not being difficulties of the sort we point out in distinguishing different such theories empirically. Consider, as an example, the case of temperature, whose relation to microphysical events is sometimes claimed analogous to consciousness' relation to physics. In contrast to consciousness, however, experiments on temperature explore a purely macroscopic theory – thermodynamics – which does not address microphysics at all. The relation between temperature and microphysics is addressed only in terms of theory reduction of thermodynamics to statisctical physics (Dizadji-Bahmani, Frigg, & Hartmann, 2010). What is more, in statistical physics, the microphysical state actually depends on temperature, as apparent for example from the fact that temperature is part of the partition function that describes the state's statistical properties (Landau & Lifshitz, 1980). If one were to change one's theory of how temperature supervenes on the physical, one would have to change these statistical properties as well so as to ensure the link to thermodynamics remains valid. Different theories of temperature are not compatible with one and the same microphysical distribution.

the basis of these experimental results. In either case, our results show that the closure of the physical must be abandoned in both theory and experiment. Theories of consciousness must explicitly state how what they take to be consciousness (physical or otherwise) comes to determine reports and other measures of consciousness, and to do this they must enter the realm of natural science.

In a very different context, Einstein once asserted that "[it] is the theory which decides what we can observe" (Filk, 2016; Heisenberg, 1971). It seems that this point has not yet been fully recognized in the construction of scientific theories of consciousness.

Part III.

On Methodology

7. Towards a Structural Turn in Consciousness Science

Johannes Kleiner[1]

7.1. Introduction

So far, the scientific study of consciousness has mainly employed verbal and linguistic tools, as well as simple formalisations thereof, to describe conscious experiences. Typical examples are the distinction between 'being conscious' and 'not being conscious', between whether a subject is 'perceiving a stimulus consciously' or not, between whether a subject is 'experiencing a particular quale' rather than another, or more generally any account of whether some X is part of the phenomenal character of a subject's experience—part of what it is like to be the subject, that is—at some point of time. Formalisations of these verbal descriptions mostly make use of set theory, examples being sets of states of consciousness of a subject and simple binary classifications, or of real numbers, for example to model 'how conscious' a system is. There are sophisticated mathematical techniques in the field, but to a large extent, they only concern the statistical analysis of empirical data, and the formulation of a theory of consciousness itself, but not the description of conscious experiences which underlies the data collection or modelling effort.

Much like words shape thoughts, descriptions shape science. In the case of consciousness studies, the descriptions that were available so far have fed into theories of

[1]Published as: Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, 119, 103653. (Kleiner, 2024)

consciousness, have determined what can be inferred about the state of consciousness of a subject, and have guided ways of conceptualising the problem under investigation.

They have, for example, led to a number of theories that explain what it takes for a single stimulus or a single piece of information to be consciously experienced, but which remain silent or vague on how the phenomenal character as a whole is determined. They have led to measures of consciousness which are specifically tailored to find out whether a single stimulus or single quality is experienced consciously (Irvine, 2013), but are not meant to infer phenomenal character beyond this. And to some extent, at least, they have privileged research programmes which search for either-or conditions related to consciousness, such as arguably the search for Neural Correlates of Consciousness (NCCs) that is largely predicated on a conception of having "any one specific conscious percept" (C. Koch et al., 2016).

Because verbal descriptions only parse part of the phenomenal character of an experience, part of what it is like for an organism to live through a particular moment, it is no surprise that means to go beyond these simple descriptions are highly sought after.

In recent years, the idea of using mathematical spaces, or mathematical structure more generally,[2] to go beyond verbal descriptions and simple formalisations have started to sprout in virtually every discipline involved in the scientific quest to understand consciousness. Following rich developments in psychophysics over more than a century (Pashler & Wixted, 2004), and pioneering work by Austen Clark (A. Clark, 1993) and David Rosenthal (Rosenthal, 1991) in consciousness science, mathematical spaces are now applied in philosophy, (A. Clark, 2000; Coninx, 2022; Fortier-Davy & Millière, 2020; Gert, 2017; Lee, 2021, 2022; Rosenthal, 2010, 2015, 2016; Fink et al., 2021; Lyre, 2022; Kob, 2023; Renero, 2014; Prentner, 2019; Yoshimi, 2007; Chalmers & McQueen, 2022; Silva, 2023; Atmanspacher, 2020), neuroscience (Tononi, 2015; Tallon-Baudry, 2022; Zaidi et al., 2013; Lau, Michel, LeDoux, & Fleming, 2022; Malach, 2021; A. Haun & Tononi, 2019; Oizumi et al., 2014; Hebart, Zheng, Pereira, & Baker, 2020; Josephs, Hebart, & Konkle, 2023; Tsuchiya et al., 2023; Zeleznikow-Johnston, Aizawa, Yamada, & Tsuchiya, 2023; Haynes, 2009; Michel, in press), cognitive science (Hoffman, Prakash, & Prentner, 2023; Rudrauf et al., 2017; Hoffman & Prakash, 2014; O'Brien & Opie, 1999), psychology (Klincewicz, 2011; Kostic, 2012; Young, Keller, & Rosenthal, 2014) and mathematical consciousness science (Grindrod, 2018; Kleiner, 2020b; Stanley, 1999; Resende, 2022; Mason, 2013, 2021; Signorelli, Wang, & Coecke, 2021; Tsuchiya, Taguchi, & Saigo, 2016; Tsuchiya & Saigo, 2021; Tsuchiya et al., 2022; Kleiner, 2020a; Kleiner & Hoel, 2021; Kleiner & Ludwig, 2024). They are known under various different names, including quality spaces (A. Clark, 1993; Rosenthal, 2015), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021; Rosenthal, 2010), qualia struc-

[2] The term *mathematical structure*, which I will explain in detail Section 7.3 below, is more general than the term *mathematical space*. That is, every mathematical space is a mathematical structure, but there are also mathematical structures which are not mathematical spaces, either because they only comprise individuals (so do not satisfy the intuition that a space is about many individuals), or because their structure is more complex than one would typically take a space to be. The question of which mathematical structures to call mathematical spaces is a matter of convention, which is why there is no definition of a general concept of mathematical space in mathematical logic.

ture (Kawakita, Zeleznikow-Johnston, Tsuchiya, & Oizumi, 2023; Kawakita, Zeleznikow-Johnston, Takeda, et al., 2023; Tsuchiya et al., 2022), Q-spaces (Chalmers & McQueen, 2022; Lyre, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink et al., 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). A first formalised theory of consciousness to make use of mathematical spaces was Integrated Information Theory (IIT) 2.0 (Tononi, 2008); more recent versions expand and refine the idea (Oizumi et al., 2014; Albantakis et al., 2023).

What unites all these proposals is the hope that the mathematical structures they propose are useful to describe the phenomenal character of an experience more comprehensively, more precisely, or more holistically than verbal descriptions or simple formalisations allow, and that mathematical structures can cope both with the apparent richness and with the many details that make up experiences. If this hope pans out, it has far-reaching implications on how to study, measure and think about consciousness.

My goal here is to offer three comments which I think are important to keep in mind when applying structural ideas in theory and experimental practice, so as to avoid misconceptions or misunderstanding early on. I hope that my comments are helpful for those working on structural ideas as well as those observing these developments with a degree of scepticism.

7.2. Three Promises of a Structural Turn

Before offering my comments below, I will briefly sketch the implications and limitations that structural methodologies may have for consciousness science. This might be of interest to those who have not engaged with this research before, and allows me to illustrate what I think are some of the great promises of a structural turn.

7.2.1. Theories of Consciousness

We currently have at least 39 theories of consciousness,[3] with new theories being proposed on a regular basis, albeit without much general attention. The reason for that, I contend, is that as far as theoretical work is concerned, it is actually very easy to come up with theories of consciousness of the type we have today.

The majority of contemporary theories of consciousness aim to explain whether a system's state, a stimulus, a piece of information, or a representation is consciously experienced, or not. That is, they target a *binary classification* between states, signals, stimuli or representations. The simple verbal distinctions mentioned in the introduction—a system 'being conscious' or not, 'perceiving a stimulus consciously' or not, 'experiencing a particular quale' or not—are all examples of such binary classifications.

Formulating theories of consciousness that target binary classification is relatively straightforward, as far as theoretical work is concerned. This is because devising a $\{0, 1\}$ classification only requires identifying some property, function, or dynamical mode of a brain mechanism. All configurations that exhibit this property, function or dynamical mode are mapped to 1, while all which do not are mapped to 0. And within non-structural approaches, nothing technical prohibits one from postulating that the 1 cases correspond to conscious experience of a stimulus, state, piece of information or representation, while the 0 cases correspond to unconscious experience thereof. The empirical or conceptual validity of such a choice is an important question, yet from a technical standpoint, formulating theories that target these distinctions is straightforward.

It is much more difficult to come up with a well-formed hypothesis that relates to a mathematical space or mathematical structure. That is because a mathematical space or mathematical structure has two parts. On the one hand, it contains a set of points. On the other hand, it contains relations or functions that express connections between the points, for example an order relation or a metric function. Therefore, there is much more information to provide when specifying how a space or structure relates to a brain

[3] An unpublished list compiled by Dr. Jonathan Mason on behalf of the *Association for Mathematical Consciousness Science* (AMCS) and the *Oxford Mathematics of Consciousness and Applications Network* (OMCAN) comprises the following theories of consciousness in the peer-reviewed literature: Activation/Information/Mode-Synthesis Hypothesis, Adaptive Resonance Theory, Attention Schema Theory, Centrencephalic Proposal, Conscious Agent Networks, Conscious Turing Machine, Consciousness Electromagnetic Information Field Theory, Consciousness State Space Model, Cross-Order Integration Theory, Dendrite/Apical Dendrite Theory, Dynamical Core Theory, Electromagnetic Field Hypothesis, Enactive and Radical Embodiment, Expected Float Entropy Minimisation, First Order Representational Theory, Free Energy Principle Projective Consciousness Model, Global Neuronal Workspace Theory, Global Workspace Theory, Higher-Order Thought Theory, Integrated Information Theory, Integrated World Modeling Theory, Layered Reference Model of the Brain, Memory Consciousness and Temporality Theory, Mesocircuit Hypothesis, Multiple Draft Model, Network Inhibition Hypothesis, Neural Darwinism Theory, Orchestrated Objective Reduction, Passive Frame Theory, Predictive Processing and Interoception, Proto-Consciousness Induced Quantum Collapse, Psychological Theory of Consciousness, Radical Plasticity Thesis, Recurrent Processing Theory, Self Comes to Mind Theory, Semantic Pointer Competition Theory, Single Particle Consciousness Hypothesis, Temporo-Spatial Theory of Consciousness, Thalamo-Cortical Loops and Sensorimotor Couplings. This list might not be complete, and some of the theories might point to similar or analogous theoretical constructs.

mechanism, or a physical system more generally. Furthermore, virtually every mathematical object comes with a set of axioms that parts of the object have to satisfy. So not only is more information needed, but this information may also have to satisfy constraints to provide a legitimate definition. This is why defining a space or structure is much more of a challenge than finding a binary classification.

The task is more difficult even if the space or structure that a theory is to provide has a specific, theory-independent form. That is the case if the theory has to account for phenomenal structure that has independent justification or independent motivation, for example from psychophysical experiments. This difficulty is illustrated by the fact that we do not, at present, have a theory of consciousness that targets the mathematical structures that have been proposed to account for conscious experiences on independent grounds. To the best of my knowledge, there are only two theories that define phenomenal spaces: Integrated Information Theory (IIT) (Albantakis et al., 2023) and Expected Float Entropy Minimisation Theory (EFE) (Mason, 2021). While both theories represent significant advances, establishing a link to existing phenomenal spaces (cf. Section 7.5) remains a next-level challenge.[4]

Because formulating theories that account for phenomenal structure in addition to non-structural explananda necessitates meeting more constraints than formulating non-structural theories, structural theories are likely to be more predictive than their non-structural counterparts. Furthermore, because the phenomenal structure is an integral aspect of phenomenal character, a theory that accounts for phenomenal structure in addition to non-structural explananda has a broader explanatory scope than one that focuses solely on the conscious-unconscious distinction. Therefore, a structural turn might deliver more explanatory and more predictive theories of consciousness. This is the first major implication I can see of structural approaches in consciousness science.[5]

Structural methodologies might inspire, and be inspired by, novel theoretical ideas that derive from any of the existing theories of consciousness, or from their combination. Proposals like the Conscious Turing Machine (L. Blum & Blum, 2022) or Integrated World Modeling Theory (Safron, 2022) that combine features of existing theories of consciousness (such as, for example, Integrated Information Theory, Global Neuronal Workspace Theory, and Free Energy Principle based proposals) could be particularly interesting in this regard.

[4] Proponents of both theories are fully aware of this task, and IIT has made a first step in this direction in (A. Haun & Tononi, 2019). In addition to accounting for phenomenal structure that has independent justification, there are other tasks and challenges that structural theories have to meet and resolve. For example, an anonymous reviewer has kindly pointed me to the fact that according to IIT, richly structured experience can be entailed by static systems without dynamics, which might pose an empirical or conceptual challenge for IIT.

[5] In saying this, I do not intend to diminish the value of "binary" theories of consciousness. They are an integral part of consciousness science and encapsulate a substantial body of evidence. But, on my view, they need to be extended so as to address phenomenal character more holistically as well. Whether this should be done on a case-by-case basis, or whether there might be a theory of qualitative character that can serve for a larger number of binary theories, is not something that needs to be decided in advance.

7.2.2. Experimental Investigations

A shift towards structural methodologies could also have significant implications for experimental research. One immediate implication follows from the previous section, i.e., from the transformative effect that structural methodologies could have on theories of consciousness. If structural theories of consciousness would indeed be more predictive than the non-structural theories we have today, then they might be easier to test than the theories we have today,[6] and the new predictions about structural facts might offer new avenues for experimental investigation.[7]

But structural thinking could also yield new experimental tools and methodologies that are separate from theoretical advancements. For instance, under certain conditions, structural approaches offer an entirely new methodology for measuring NCCs (Fink et al., 2021). This methodology could potentially address some of the foundational challenges in existing methodologies, such as the co-activation of cognitive processing centres causally downstream of the core NCC, and might not require traditional methods to assess a subject's state of consciousness. I discuss and criticise the key assumption that enables this methodology—the assumption of a structure-preserving mapping between phenomenal and neuronal structures—in Section 7.4 below. But nevertheless, even if this assumption proves to be more limited in scope or strength than initially anticipated, the methodology might still have advantages compared to existing options to search for NCCs.

The implication that intrigues me most, however, is the possibility that structural approaches may introduce new *measures of consciousness*. A measure of consciousness, as conventionally understood, is a method to determine whether an organism is conscious, or whether a given stimulus or signal has been consciously perceived. Measures of consciousness are "consciousness detection procedures" (Michel, 2023) of sorts.

Building on the extensive previous work in both psychophysics and consciousness science, structural approaches raise the possibility to construct new and potentially more powerful measures of consciousness, which do not only focus on whether a single stimulus is experienced—a single quality of phenomenal character, that is—, but on phenomenal character more comprehensively.

The potential of structuralist approaches in this regard can be nicely illustrated by considering verbal report, which is a paradigmatic (albeit often criticised) measure of consciousness. In the case of report, subjects use language to report facts about their experience. They might, for example, indicate that they experienced a red colour, or saw a face in a masked stimulus. The problem with reports is that when compared with the actual experience, they contain very little information. Which shade of red did the subject

[6] Lukas Kob made this point for *structuralist* approaches during a wonderful talk at the recent *Structuralism in Consciousness Studies* workshop at the Charité Berlin, though my comment here concerns the wider scope of *structural* approaches, cf. Section 7.3 for more on that distinction.

[7] Speculating wildly, one might hope that if theories of consciousness could account for *theory-independent* phenomenal spaces, this could help to mitigate the problem that empirical tests of theories of consciousness currently rely heavily on theory-dependent *methodological choices* (Yaron, Melloni, Pitts, & Mudrik, 2022).

experience, precisely? How did they experience the face, and with which details? What else did they experience in addition to the reported fact? In information-theoretic terms, this problem arises because the channel capacity of verbal report and other behavioural indicators is low compared to the information content of conscious experiences.[8]

Structural approaches allow us to bypass the limited channel capacity of reports and similar measures of consciousness, because structural descriptions can *store information* about the phenomenal character of a subject. That is the case because structural descriptions represent features of a subject's phenomenal character that relate individual non-structural facts. For example, relations between experiences, or relations between constituents of experiences, such as individual qualities.

Given the structural information in a phenomenal space, a few bits of information collected in an experimental trial, for example by means of reports or similar measures of consciousness, can suffice to pin down the location in a structure, resulting in information about what a subject is experiencing that might go far beyond the bits of information that were collected. This is similar to how a geographic map can be used to decode rich information about one's location based on a few bits of information. Finding one's way in the wilderness without a map or map-like tools generally is a very difficult task. But given a map, procedures like triangulation are available that only require a few bits of information, such as the angles between three landmarks in line of sight, to pin down one's position and find one's way. That is possible because maps store information about geography. Another example of this sort is quantum tomography, where a set of carefully chosen measurements, together with structural information about the quantum state (specifically, the inner product and projective structure of the Hilbert space) is used to pin down the exact state among an infinite number of possibilities.

In a similar vein, phenomenal spaces might be used to decode information from carefully chosen low-channel-capacity measures of consciousness. How to precisely do this remains an open question as of yet, and strongly depends on a thorough understanding of phenomenal structure in the first place (cf. Section 7.5), but it is a viable possibility.

7.2.3. Conceptual Work

Structural approaches can also be essential, finally, in *conceptualising consciousness and its potential problems*. It is not unlikely that interesting philosophical implications arise, specifically in the context of structuralist assumptions, but what I'd like to highlight here is the importance of structural thinking in shaping our pre-theoretic problem intuitions about consciousness; those intuitions, that is, which guide both our theorising and experimental work.

Structural thinking might well turn what we previously thought about consciousness upside down. It might change how many of us think about their own research in the

[8] I'm very grateful for a conversation with Lucia Melloni about the problems of reports and structural ideas to resolve these during a walk at the above-mentioned *Structuralism in Consciousness Studies* workshop. The idea sketched here came up during this walk, and is Lucia's as much as, or even more than, mine.

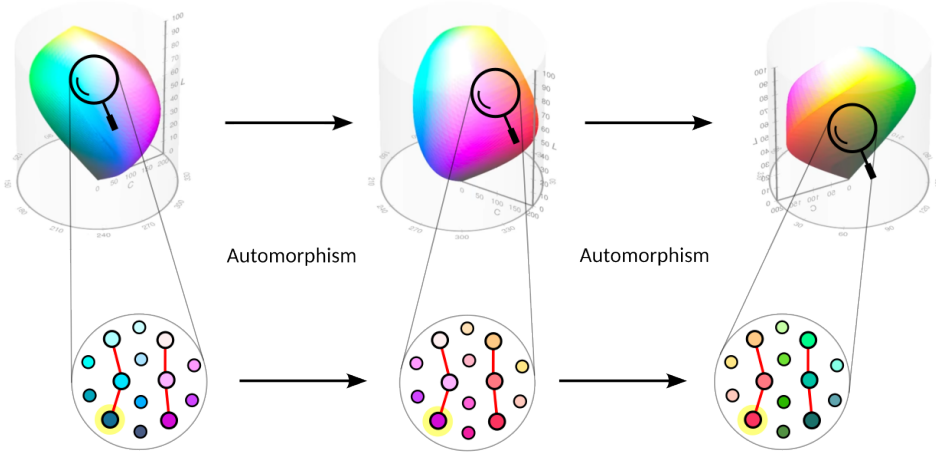


Figure 7.2.1.: **What is an automorphism?** This figure illustrates the concept of automorphisms. Automorphisms are somewhat analogous to rotations of a space around some axis (top row). More formally, an automorphism is a function that maps every point of a mathematical space to a different point of the same space in such a way that all relations of the space are preserved: whenever two points are related before the mapping, they are also related after the mapping. This is illustrated by the bottom row, where individual points of the space are depicted by coloured dots, and relations are depicted by red lines. An automorphism maps every dot to a new dot, represented here by the change in location of the colours, in such a way that when two dots were related before the mapping (red line between two dots) the targets of the mapping are also related (red line between target dots). Automorphisms form a group because any two automorphisms can be combined to form another automorphism, in this case one from the left-hand space all the way to the right-hand space.

(Depiction of CIE colour space gamuts by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image is shared under the same license.)

first place. To give two very preliminary examples, I think that structural approaches are relevant for epistemic arguments like Mary's room (Jackson, 1998, 1986), and for modal arguments like colour inversion (Shoemaker, 1982; Block, 1990).

For epistemic arguments such as Mary's room, the big question is whether one presumes that structural facts about experiences are known. If Mary propositionally knows, for example, which structure the experience of red has, and if structure is sufficient to individuate experiences, then she might be able to use her advanced neuroscience knowledge to create an embedding of the structure of red experiences within her own phenomenal space, even if she never experienced red, or any colour for that matter, before. Similarly, outside the realm of thought experiments, we might use structural facts to create experiences that approximate what it is like to be a bat. Structure might furnish an objective phenomenology (Lee, 2022).

Modal arguments, similarly, need to be rethought. The typical colour inversion thought experiment presumes fairly homogeneous colour spaces—colour spaces that possess symmetries. This presumption is critical because if a colour inversion is not a symmetry, then the difference between colour experience before and after the inversion will manifest itself both in behaviour and in the use of colour words: through similarity judgements and other expressions of structural facts. The closest approximation we have to a space of consciously experienced colour qualities is the CIELAB colour space (Schanda, 2007), a rendering of which is depicted in Figures 7.2.1, 7.3.2, and 7.5.1, which is highly non-homogeneous and may not admit symmetries to the extent that we expect. Adding valence and other consciously experienced attributes of colour experiences might further erode any remaining symmetries. Thus, at least the usual intuitions regarding qualia inversions and other modal arguments may cease to be valid. Structural approaches might force us to reconsider intuitions that are built on these types of arguments.

7.2.4. Limitations

While structural approaches do, on my view, offer a number of benefits to the science of consciousness, it is also important to see their limitations.

A first limitation of structural approaches is that it is not clear, at present, how much of phenomenal character—how much of what it is like to experience something, that is—can be grasped by structural tools. While it is clear that much of the phenomenal structure that is usually associated with the content of consciousness can be represented structurally (much of it actually *is* structural, one might say), it is not clear whether some of the more subtle or remote facets of phenomenal character are amenable to a structural analysis. Can the experience of a self or ego be represented structurally? What about the experience of other minds? Or the pre-reflective and pre-conceptual awareness of being aware, sometimes referred to as subjective character?

A second limitation of structural approaches relates to measurability. Even if a facet of phenomenal character is amenable to structural tools, it might still be difficult, costly, or even impossible to measure. It might take years to construct a full quality space of

a single modality. Is this actually feasible in experimental practise for anything but the most salient structures of phenomenal character?

A third limitation is the question of whether structural approaches can actually get any closer to modelling what is sometimes described as an intrinsic nature of qualia or qualities, the "raw experience", so to speak. Do structural approaches have any handle on modelling this? Or can they just circumscribe the structure that intrinsic properties instantiate? And to the extent that such intrinsic nature is the core of the problem of consciousness, can structural approaches get us any closer to understanding this core?

My own view of these limitations is that they define some of the key questions that structural research will have to tackle in the upcoming years. Because experiences exhibit structure, structural approaches are, by necessity, part of any research programme that targets experiences in full. But to what extent they contribute to resolving the core questions at the heart of consciousness science is an open question.

7.3. Metaphysical Premises

My first comment concerns an intuition which I have often encountered when discussing structural approaches with colleagues: that structural approaches are metaphysically presuming. Most notably, they seem to many to be tied to physicalist or reductionist metaphysics. The goal of this comment is to show that this is not the case. Structural approaches offer a new descriptive tool that can—in theory, at least—be applied independently of metaphysical assumptions, and in research programs of any metaphysical flavour. Structural approaches do not in themselves have metaphysical premises, and they do not by themselves come with a preferred metaphysical interpretation. Rather, they can be applied to and combined with the particular metaphysical ideas or presumptions that are already employed in a research program.

The major reason why structural approaches are often taken to be metaphysically presuming is that they are conflated with structuralist approaches. Structuralist approaches assume that individuals can be individuated by structure: that for every individual x, there is a unique location in a structure, a location in which only x holds. Intuitively speaking, the idea is that specification of all structural facts suffices to also specify all facts about individuals in that structure.

In the context of consciousness science, the individuals in question can be experiences, phenomenal character, qualities or qualia. The structures in question are experience spaces (spaces whose elements are experiences), phenomenal spaces, quality spaces or qualia spaces. Furthermore, there are ontological, epistemological and methodological ways of reading a structuralist claim. But in all cases, the idea is that the domain of individuals exhibits structure, and that this structure is sufficient to individuate the individuals in the relevant sense.

Structural approaches, in contrast, are not committed to a claim of individuation. An approach is structural if it applies mathematical structure. And as I will now explain, more often than not, mathematical structure does not individuate individuals. In order

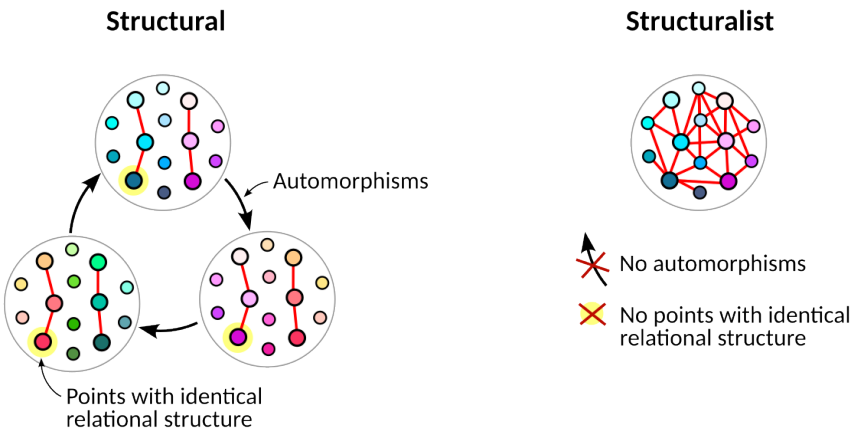


Figure 7.3.1.: **Structural vs. structuralist approaches.** Structural approaches make use of mathematical spaces or mathematical structures to represent or describe conscious experiences. These spaces and structures may, and in general do, admit for automorphisms (cf. Figure 7.2.1). This implies that there are points in the space which have the exact same relational structure. Structuralist approaches, on the other hand, assume that all points of the space can be individuated by their relational structure, meaning that no two points have the same relational structure. This can only hold true if the space does not admit automorphisms, other than the identity mapping that is always an automorphism.

to see why, we must differentiate between two readings of the term 'structure'. This will also yield a clear formal definition of structuralism in a given consciousness-related domain.

Mathematics offers an unambiguous definition of what a structure is. A *mathematical structure* consists of two things: domains, on the one hand, and functions or relations, on the other hand. The *domains* of a structure are the sets on which the structure is built. They comprise the points, or elements, in a space, the individuals in a structuralist sense. In the case of a metric space, for example, there are two domains: the set of points of the metric space and the real numbers that constitute the "distances" between points. In the case of a partial order, there is just one domain: the domain of elements that are to be ordered. The second ingredient of a mathematical structure are functions and/or relations. *Functions* map some of the domains to other domains. In the case of a metric structure, for example, there is a metric function that maps two points to a real number. *Relations* link points to each other. In the case of a partial order, for example, there is a binary relation on the set of points. This relation specifies ordered pairs of points, usually written as $p_1 \leq p_2$.

When the term 'structure' is used in natural science, it usually follows this mathematical definition. For example, if we talk about the structure of space-time, we mean the mathematical structure that describes space-time, called a Pseudo-Riemannian manifold. If we talk about the structure of a neural network, we mean the mathematical structure of the directed graph that specifies the connectivity of the network: the mesh of nodes and edges, where each node represents a neuron or neuronal assembly, and where each directed edge specifies a neural pathway between neurons or assemblies.

When we use the term 'structure' in the context of structuralist ideas, however, it only refers to the second ingredient of a mathematical structure: the functions and relations that a mathematical structure contains. These functions or relations are what individuates the individuals—the elements of a domain—in a structuralist sense.

While customary in the context of structuralist assumptions, this use of the term 'structure' to designate only relations and functions is problematic. That is the case because we cannot actually specify relations or functions without specifying the points or elements that the relations or functions operate on. The symbol '$\leq$', for example, can be used to indicate a type of structure, a partial order in this case, but it cannot define or specify a structure. Any concrete definition or specification of a partial order needs to make use of, or refer to, the points that the relation links. It needs to make use of some set of points—some domain in the mathematical sense of this word. Strictly speaking, it does not make sense to use the term 'structure' to refer *only* to the functions or relations. I will refer to structure in the structuralist sense—that is to the functions and relations that are part of structure in the proper sense of the term—as *structure in the narrow sense of the term*.

The structuralist idea that relations or functions determine all individuals still makes sense, of course, independently of terminological issues. And it can be expressed in a neat formal requirement, making use of the notion of an automorphism, cf. Figure 7.2.1. An automorphism is a one-to-one mapping from the domains of the structure to them-

selves which preserves the functions or relations. That is, it preserves structure in the narrow sense of the term. For every point of the structure, an automorphism specifies a point as its target in such a way that the functions and relations of the structure do not change when going from the source to the target: whenever some points satisfy a relation before the mapping, they also satisfy the relation after the mapping, and equally so for functions.

Automorphisms may or may not exist. The identity mapping (not changing anything) is always an automorphism, but depending on how rich or complex the structure in the narrow sense of the term is, there might not be other automorphisms. In particular, if it is indeed the case that every point x of a structure satisfies a unique location of structure in the narrow sense of the term, then there is no automorphism other than the identity. One cannot exchange any two points without changing structure in the narrow sense of the term. In this case, one says that the *automorphism group is trivial*. [9] Vice versa, if the automorphism group of a structure is trivial, then every point must have a unique location.[10]

Because structuralism (in the context of consciousness) is the assumption that every point x of a structure (in the general sense of the term) satisfies a unique location of the structure in the narrow sense of the term, structuralism is equivalent to the condition that the automorphism group of the relevant structure (in the general sense of the term) is trivial. This constitutes a nice formal characterisation of structuralism in consciousness science:

> **(STR)** Structuralism about a domain is true iff the automorphism group of that domain is trivial.

Here, the domain could comprise individual experiences, phenomenal characters, qualities or qualia, depending on which type of structuralism is under consideration.[11]

The crucial point of this section is that mathematical structures can, but need not, obey structuralist assumptions; they may or may not have a trivial automorphism group. A theory or experiment can be *structural*, in the sense that it makes use of mathematical spaces or structures, without necessarily being structuralist. This is illustrated by Figure 7.3.1.

In fact, if we look at mathematical spaces in mathematics, physics and other natural sciences, in the majority of cases, the automorphism group is *not* trivial. Simple examples of spaces with non-trivial automorphism groups are the Euclidean spaces $\mathbb{R}^2$, $\mathbb{R}^3$ and $\mathbb{R}^n$ for any $n \geq 2$, and many metric spaces, Riemannian manifolds, Hilbert spaces, or graphs.

[9] It's 'trivial' because that's the simplest possible case, and the set of automorphisms is a group because automorphisms can be combined and inverted as required by the axioms of a group in mathematics.

[10] For every point to have a unique location in a structure is for there not to exist a permutation or other mapping of the domains of that structure to themselves that leaves the structure in the narrow sense of the term invariant.

[11] The term 'domain' also has two meanings: The meaning of domain in the sense of mathematical structure as introduced above, and the meaning of domain as a group of related items in general language. Fortunately, both work if it is clear what the structure of a domain is.

Therefore, not only is there a difference between structural and structuralist approaches, but it is in fact quite common that the former applies while the latter doesn't. Structures in the general sense of the term may, but often do not, boil down to structures in the narrow sense of the term. This has three consequences for research in a structural turn.

Consequence 1. Structural vs. Structuralist Agendas

Much like the two senses of the term 'structure' at issue here are often conflated, so are structural and structuralist agendas. Both are subsumed under the general heading of 'structuralism', for example. A first consequence of the above is that there is a difference between structural and structuralist agendas, and it is important to be clear about which agenda one is pursuing when engaging in structuralist research.

If one is using mathematical tools and methods, for example, to help place "structural phenomenal properties at the core of the science of consciousness" (Chalmers, 2023b), as required by a very attractive position called weak methodological structuralism that has recently been put forward by David Chalmers, then one is engaging in a structural agenda: an agenda which makes use of mathematical spaces and mathematical structure but which is not committed to a structuralist claim. Put differently, structural tools like mathematical spaces can also be employed if one rejects the idea that structure (understood in the narrow sense of the term) is all that matters. They are free of explanatory and epistemic charge.

Consequence 2. Metaphysics of the Mind

Many structuralist approaches are not metaphysically neutral. They imply that certain properties that some consider crucial with respect to consciousness do not exist, or are not knowable. For example, if ontic phenomenal structuralism is true, then there are no intrinsic phenomenal properties, and no genuinely private properties. Ontic structural realism implies that there are no qualia as conventionally understood (Dennett, 1988). If epistemic phenomenal structuralism is true, then one cannot know (either scientifically, or by introspection) of intrinsic or private properties, all we know about conscious experiences derives from structural properties.

Structural approaches are not tied to these assumptions. They are perfectly compatible with the existence of intrinsic or private properties. As far as the mathematics is concerned, if private or intrinsic properties exist (or if there are properties which are not accessible to structural cognitive processing), this simply means that the automorphism group of the structure is not trivial. There are points that cannot be individuated by structure alone.

To give a very simple example, consider the case where there is no structure in the narrow sense of the term at all, i.e. the case where there are no relations or functions between qualities or qualia at all. This case can be described in terms of mathematics: the qualities or qualia simply form a set. A set is a mathematical structure according

to the definition of mathematical structure referenced above. It is the simplest case of a mathematical structure, but an important one. So while this case is opposed to the ideals of structuralist thinking, it is a simple but perfectly fine example of a structural approach.

What is more, structural approaches might actually help to address intrinsic, private or ineffable properties in scientific contexts. My first paper on consciousness, (Kleiner, 2020b), is devoted precisely to this issue. In a nutshell, I show that mathematical tools can be used to formulate theories of consciousness that address these properties even if they are, in an intersubjective sense, non-collatable. Because of these mathematical tools, mathematical approaches allow us to go further than non-mathematical approaches can go. Ultimately, this works because, in the words of Jürgen Jost, "[m]athematics translates concepts into formalisms and applies those formalisms to derive insights that are usually not amenable to a less formal analysis." (Jost, 2015).

Consequence 3. Metaphysics beyond the Mind

The third consequence, finally, concerns the conviction mentioned at the beginning of this section that structural approaches seem to many to be tied to physicalist or reductionist metaphysics.

The intuition that motivates this conviction arguably derives from the equivocation of structural and structuralist assumptions, together with the idea that science can only explain relations. If structural assumptions would indeed imply that "[t]here is nothing to specifying what something is over and above stating its location in a structure" (Fink et al., 2021), and the physical sciences could only explain structure, then it would indeed be the case that structural approaches would render consciousness amenable to scientific and arguably physicalist explanation. What is more, when ontology is concerned, structuralist assumptions imply that none of the prototypical non-physicalist properties of consciousness exist (cf. Consequence 2). This, too, intuitively speaks in favour of a physicalist and reductionist research programme.

While it is clear that these intuitions do not have the force of a logical argument, it seems fair to say that *structuralist* assumptions are well aligned with physicalist metaphysics, and in the form of one of its most promising incarnations, neuro-phenomenal structuralism (Fink et al., 2021; Lyre, 2022), might even "open an attractive door for reductionism" (Fink et al., 2021).

The problem with the conviction mentioned above is that structural approaches are not necessarily structuralist approaches. The majority of mathematical spaces that are used in the sciences have a non-trivial automorphism group and therefore do not satisfy the defining criterion of a structuralist approach in the context of consciousness science (cf. Figure 7.3.1). In other words, one can choose to apply mathematical tools and methods to describe consciousness without committing to structural assumptions and a fortiori without committing to physicalist or reductionist metaphysics. Structural approaches can be used and might be beneficial in any type of metaphysical programme, from reductive physicalism to property dualism or idealism.

In fact, there are a number of structuralist approaches which target non-physicalist metaphysics already, on the level of toy-models. (Atmanspacher, 2020), for example, uses mathematical language to outline how the neutral domain in a Pauli-Jung style dual aspect monism might relate to the mental and physical aspects. And the proposals (Signorelli, Wang, & Coecke, 2021) and (Signorelli, Wang, & Khan, 2021) use a category-based graphical calculus to indicate how ideas from the Yogacara school of Buddhist philosophy could be fleshed out in terms of a scientific theory of consciousness.

I make these points not to argue for a non-physicalist research programme, but to show that structural approaches are not tied to physicalist or reductionist assumptions. Rather, my point is that mathematical spaces and mathematical structures provide descriptive tools that can be applied to any choice of metaphysical assumptions, and in research programmes of any metaphysical flavour. Structural approaches do not have metaphysical premises, and they do not come with a preferred metaphysical interpretation.

7.4. Isomorphisms and Structure-Preserving Mappings

The core question which drives the scientific study of consciousness is the question of how conscious experiences and "the physical" relate. A ubiquitous mathematical object in the context of mathematical structures is that of an *isomorphism*, illustrated in Figure 7.3.2 and explained in detail below. Because of its ubiquity, when introducing structure to the phenomenal domain, many feel that it is natural to assume that this structure is related to physical structure by an isomorphism or structure-preserving mapping. My goal here is to show that this assumption is not in fact justified. We either need to search for a rigorous justification, or if there is none, proceed in different ways.

Intuitively speaking, an isomorphism expresses a relation between two structures. Precisely speaking, it is a bijective mapping *between the domains* of two structures that preserves the relations or functions of these structures. That is, it is a map from the elements or points of one structure to the elements or points of another structure. A map is bijective if it is one-to-one and onto, meaning that every element in the target space gets mapped to by exactly one element in the source space.

In practice, because the physical has a much larger domain and much richer structure than the phenomenal, when the concept of an isomorphism is applied in consciousness science, what is actually meant is an *isomorphism onto the image*. This means that there is an isomorphism from the phenomenal domain to a substructure of the physical domain. Often, homomorphisms are used as well. They are defined exactly like isomorphisms, except that they do not have to be one-to-one or onto, so that some elements in the target space might not get mapped to, and/or several elements in the source space might map to the same element in the target space. Strictly speaking, though, homomorphisms are not appropriate either,[12] but to avoid unnecessary technical details, I will

[12] For the nerds ;-) The concept of homomorphism as used in mathematics presumes that two structures

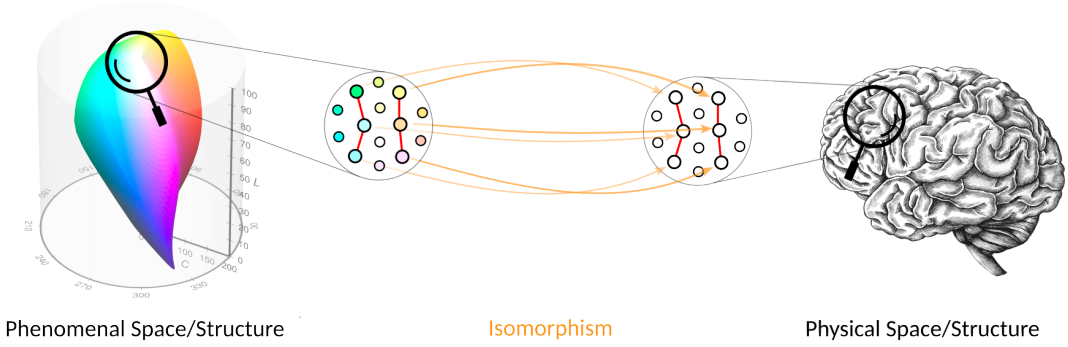


Figure 7.3.2.: **What is an isomorphism?** This figure illustrates the concept of isomorphisms as applied in consciousness science to link a phenomenal space or structure (left) with a physical space or structure (right). By definition, isomorphisms operate on the level of points. An isomorphism maps every point of the phenomenal space to a point in the physical space. It does so in such a way that the relations between points (indicated here by red lines) are preserved, meaning that any two points which are related on the left are related in the exact same way on the right, and the mapping needs to be invertible. An isomorphism presupposes that structures on both sides of the mapping are given. It does not define, or pick out, the structure in its target domain, which is why it is not a suitable mathematical object to explain, predict, or define phenomenal structure in terms of physical structure.

(Depiction of CIE colour space gamut by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image, excluding the drawing of the brain, is shared under the same license. Drawing of the human brain from Freepik.)

admit them too. I will use the term *structure-preserving mapping* to denote homomorphisms or isomorphisms with the understanding that the domains and structures of the source and target have been adapted appropriately to avoid the technical problems. As far as intuition is concerned, my comments are easiest understood when thinking about an isomorphism onto the image.

The assumption under discussion then is:

> **(ISO)** The physical and the phenomenal are related by a structure-preserving mapping from the phenomenal domain to the physical domain.[13]

This assumption is a very consequential assumption. It promises, for example, a new methodology for measuring Neural Correlates of Consciousness (NCCs). To date, NCC research has to make use of intricate measures of consciousness (Irvine, 2013), to distinguish between trials where the subject perceives a stimulus consciously from trials where it doesn't. If (ISO) is true, a whole new avenue for investigating NCCs is available: to search, among neural structures in the brain, for structures that are homomorphic to or identical with the structures of the phenomenal domain. This search could, in principle, be carried out independently of any measure of consciousness, and might give a unique result, so that potentially at least there is a methodology where one "[does] not have to worry whether subjects 'really' had a phenomenal experience of a stimulus" (Kob, 2023).

The existence of a structure-preserving mapping between the phenomenal and physical domain also has important consequences for theories of consciousness: it implies that a large class of theories of consciousness is false, namely all those which do not

---

have the same *signature*, meaning that both structures need to have the same type of functions or relations: the same number of functions or relations of the same arity, that is. Because the physical has much more structure than the phenomenal (think about the rich structure of electrodynamics in the case of neurons, say), the concept of homomorphism is too strong to express the underlying idea. One could attempt to define a *partial homomorphism* as a homomorphism that respects some, but not all, structures of the target domain. But for questions other than multiple realizability, the 'isomorphism onto the image' conception seems to be closer to the underlying intuition. The same applies if one reverses the direction of the homomorphism, cf. Footnote 13.

[13] In addition to the problem mentioned in Footnote 12, there is also the question of which direction a homomorphism should take. Should it go from the physical domain to the phenomenal domain, as in (Fink et al., 2021), or vice versa? Because it is unlikely that all elements of the physical domain are mapped to the phenomenal domain (there are neural mechanisms which are not relevant for conscious experiences, for example), and because a map in the sense of mathematics requires a specification of a target element for every element of the source domain, it seems more natural to me to choose the phenomenal-to-physical direction. Choosing the physical-to-phenomenal direction would require one to introduce yet another sense of partiality, that of a partial function, which is only defined on some of its elements. The problem with this is that a homomorphism which is partial in both this sense and the sense of Footnote 12 always exists, so that the statement (ISO) is vacuous. This is not the case for an isomorphism onto the image in the phenomenal-to-physical direction, because of the need to specify a target element in the physical for every source element in the phenomenal in such a way that the image has the same structure as the phenomenal. This is why I think isomorphisms onto the image in the phenomenal-to-physical direction are the right tool (and the right intuition) to work with, though my comments below do not turn on this choice.

take the form of a homomorphism. A good example of this is Integrated Information Theory (IIT) (Oizumi et al., 2014; Albantakis et al., 2023). It is sometimes assumed that IIT is structure-preserving or even an isomorphism, but according to IIT's mathematical formulation, this is not the case. The mathematics of IIT come with two clear 'slots' for the physical and phenomenal domain. One of the slots is the input to the theory's algorithm. It requires a physical description of a system, for example in terms of neurons. The other slot is the output of the theory's algorithm. For every system and physical state of this system, this output is a mathematical structure called 'Maximally Irreducible Conceptual Structure' in IIT 3.0, and 'Φ-structure' in IIT 4.0. This structure "is identical to [the system's] experience" (Oizumi et al., 2014). The mathematical algorithm of the theory specifies a mapping between those two slots which is not a homomorphism. Therefore, the theory does not specify a homomorphism between the physical and phenomenal domains. And consequently, if (ISO) is true then IIT must be wrong.[14]

Are isomorphisms justified?

The above shows that (ISO) is indeed a very consequential assumption. This would be good news if (ISO) were also a justified assumption. But, as I will argue here, this is not the case. While isomorphisms and homomorphisms are natural in mathematics, they appear not to be the right sort of object to achieve the goals of consciousness science in investigating the relation of the phenomenal and the physical. For the purpose of this discussion, I will assume that these goals are "to *explain*, *predict*, [or] *control* the phenomenological properties of conscious experience" (my italics) in terms of physical properties, following Anil Seth's *Real Problem of Consciousness* (Seth, 2022), with the understanding that phenomenal structure is an integral part of phenomenal character, and that structural properties are properties too.

My comments are tied directly to what an isomorphism or homomorphism is. As explained above, isomorphisms and homomorphisms are mappings between the domains of two structures (between the *points* or *elements* of these structures, that is) which satisfy certain conditions. The conditions enforce that the mappings are compatible with the structures on both ends. This has two important consequences for the question at hand.

[14] The only way to enforce viewing IIT as an isomorphism is by claiming that the output of IIT's algorithm is itself a physical structure, which then happens to be related by an isomorphism to the phenomenal domain. Given the interpretation of the mathematical structure outputted by IIT as "identical to [the system's] experience" (Oizumi et al., 2014), it is hard to see how such interpretation can plausibly be made. The mathematical quantities outputted by IIT do not appear anywhere else in the physical sciences, and are conceptually and mathematically rather removed from physical theories. Such claim also violates the implicit presupposition in (ISO) that there are more or less well-defined structures on both the phenomenal and physical sides. If there were no constraints on which structure to consider, then (ISO) would be a vacuous statement. Any mapping of the form $f : P \to E$, where P denotes physical structure and E denotes phenomenal structure, can be turned into an homomorphism between the physical and the phenomenal if E is taken to be a physical structure as well. As a rule of thumb, if a structure is actively defined by a theory of consciousness, rather than just adapted from some other part of science, it should probably not count as physical structure in the sense required by (ISO).

The first consequence is that a homomorphism *presupposes* that the structures on both ends of the mapping are given. If only one of the two structures is given, or none even, then (ISO) becomes an empty statement. This is because *any* mapping of the form $f : E \to P$, where P denotes the physical domain and E denotes the experiential domain, can be turned into a homomorphism if at most one domain comes with structure. One can simply define the structure on the other domain so that the mapping becomes a homomorphism. Assuming that there is a homomorphism without presupposing that structure on both ends of the mapping is given amounts to not assuming anything at all.

But if a homomorphism presupposes structures on both ends, it doesn't explain, predict or allow to control these structures. Homomorphisms fall short to explain, predict or allow to control those phenomenal properties they were introduced to cope with.

Second, and more importantly in my opinion, homomorphisms do not have the right mathematical form to *pick out* which structure there is. That is the case because they are maps from domains to domains. They do not actually map from structures to structures, as is sometimes thought. They only map points in one domain to points in another domain in such a way that the mapping between the points *preserves* or *respects* the structure on both ends. This speaks against an explanatory or predictive function as well, as I shall now explain.

Let us first consider explanation. Do homomorphisms, or other structure-preserving mappings, *explain* phenomenal structure in terms of physical structure? There are various notions of explanation that are available in science, ranging from the early deductive-nomological and inductive-statistical ideas studied by Carl Hempel (Hempel & Oppenheim, 1948; Hempel, 1962) to more modern understandings of explanation in the form of causal-mechanical models (Salmon, 1984), unificationist models (Friedman, 1974; Kitcher, 1989), contrastive explanation (van Fraassen, 1980) or interventionalist models (Woodward & Hitchcock, 2003; Hitchcock & Woodward, 2003).

It is clear that homomorphisms do not fit the original Hempel models of explanation because they do not derive phenomenal structure in any meaningful sense from a general law and initial conditions. What is crucial though is that they also don't sit well with the other models of explanation. This is the case because, in one form or another, these models all require 'what if things had been different' information. In the causal-mechanical model of explanation, 'what if things had been different' information is required to test the robustness of a purported causal mechanism. In unificationist models it matters for questions of breadth of a unifying explanation. In contrastive explanations it is central to deal with alternative scenarios that would have occurred under different conditions. And in interventionist models, it is required to explicate how an intervention changes the explanadum variable.

Homomorphisms do not pick out structure on the physical or phenomenal side, they only relate points of the domains in a structure-preserving way. Therefore, they do not provide 'what if things had been different' information about phenomenal structure. But 'what if things had been different' information is required by the above-mentioned models of explanations. Therefore, homomorphisms do not constitute an explanation of

phenomenal structure according to these models.

Because homomorphisms don't pick out phenomenal structure, they do not offer alternatives to how phenomenal structure could be if things had been different. For this reason, they also do not predict phenomenal structure. Prediction, too, requires mathematical tools that pick out the right structure among a class of possible structures.

A helpful way to think about the problems of explanation and prediction is to think about what would *define* phenomenal structure in terms of neural structure, or physical structure more generally. Consider, as an analogy, computer games. Computer games employ mathematical structure to model rich and detailed visual imagery. Yet the mathematical models are defined mostly in terms of objects in the sense of object-oriented programming. There is nothing in the actual code of the game which resembles the structure of the visual scene; rather, the code defines how the structure should be rendered, and it does so in terms of objects and properties. The visual structure created by the game is not homomorphic to the code that runs in order to create the scenes, yet it is defined by the code. This example illustrates that homomorphisms are not the kind of thing one would expect when defining structure.

What these points illustrate, on my view, is that homomorphisms and structure-preserving mappings more generally are not the right sort of object to define, explain, predict or control phenomenal structure. They might be natural in the context of mathematical questions, but they are not natural for the purposes of consciousness science.

Consequently, (ISO) is not in fact a natural or justified assumption. We either need to search for a rigorous justification, or if there is none, proceed in different ways. Because (ISO) is so consequential for theoretical and experimental work, using (ISO) without proper justification, or in the hope that a justification will eventually be found, is not a viable option.

This comment also applies to mathematical objects known under different names, if these objects are in fact homomorphisms. Important examples thereof are mathematical objects known as diffeomorphisms, which are maps between smooth geometric shapes called manifolds. Diffeomorphisms are homomorphisms between the mathematical structures that define smooth manifolds. And much like the simpler cases discussed above, they map points of one manifold to points on another manifold in a way that respects the mathematical structure on both sides of the map. They do not explain or define the structure.

What, if not isomorphisms?

If isomorphisms and homomorphisms are not the sort of thing that explains, predicts or defines phenomenal structure, what is? Which mathematical objects should we use to relate the physical and the phenomenal in a structural turn?

My view is that there is no general mathematical principle that we can commit to. Rather, much like theories of consciousness in the pre-structural area were built one-by-one, we have to build structural theories one-by-one, working with different ideas, concepts, motivations and metaphysics in each case. The challenge of finding the right

mathematics to explicate these ideas, concepts and motivations in a structural context is not something we can bypass by choosing one mathematical tool that fits them all. This is not technically possible, but also it is not desirable. The difference between ideas, concepts and metaphysical underpinnings in a structural context is precisely in the mathematics that relate the physical to phenomenal structure. We cannot waive the problem of finding the right mathematics without also waiving the possibility of choosing different metaphysical or conceptual ideas.

7.5. Which Phenomenal Structure?

My final comment concerns the question of which structure to consider when embarking on structural research. That's the question of what phenomenal structure *is* and how we find it. This question is important because conscious experience does not "come with" mathematical structure in any direct sense. There is nothing in what it is like to experience something that is per se mathematically structured, other than if one explicitly experiences something mathematical.[15]

Rather, mathematical spaces and mathematical structures are *tools* or *languages* we can use to describe (or model) phenomenal character, much like English or any other language can be used to describe phenomenal character. And just as we need definitions or conventions to apply English language terms, we need definitions or conventions to apply mathematical terms. These might not be as simple as in the case of English, but still they flesh out the conditions under which one is, and under which one isn't, justified in making a structural and mathematical claim.

Because mathematics is a different type of language from English, the definitions or conventions to apply structural terminology are of a different type too. They constitute *methodologies*, meaning they are collections of methods, procedures or rules, that can and need to be used to assess mathematical claims.

Because phenomenal character does not "come with" mathematical structure in any direct sense, any claim about a structural fact, and any application of structural ideas, is always *relative* to a specific understanding of what phenomenal structure is, and a fortiori, relative to the methodology that defines this particular understanding. That is, it is not meaningful to claim that experiences have a certain structure. Much like a claim about whether experiences have qualia depends on what exactly one takes the term qualia to denote, the claim that experiences have a certain structure depends on what one takes phenomenal structure to denote (Figure 7.5.1). When working with or thinking about phenomenal structure, we need to be clear about which methodology we presume. Otherwise, we're prone to making errors. This is the first major point I'd like to make in this comment.[16]

[15] We do experience mathematical structures if we know and recognize them, for example in the case of geometrical shapes, or if we actually work with mathematical structures. But we do not experience non-mathematical experiences as mathematically structured. We do not, for example, experience colours as constituting a metric space or having a partial order.

[16] Therefore, working with mathematical structure in consciousness science is different from working with

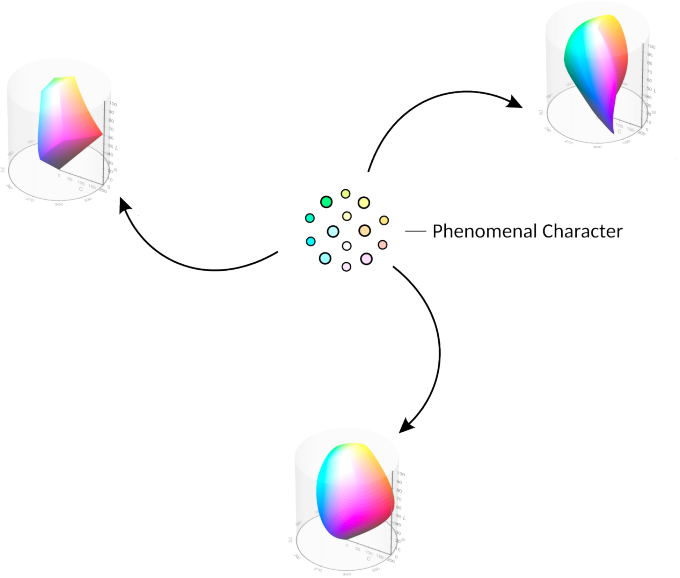


Figure 7.5.1.: **Different definitions imply different spaces.** Mathematical spaces and mathematical structures are tools to describe or represent phenomenal character, much like technical language terms are too. Different definitions or conventions of how to use mathematical terms to describe or represent phenomenal character—different conventions of what terms like 'mathematical structure of conscious experiences' or 'phenomenal spaces' mean—lead to different structural representations of the same set of experiences, here illustrated by three different CIE colour spaces. Black arrows indicate different definitions or conventions, which imply different methodologies for constructing phenomenal spaces in the lab. Much like technical language terms might differ in scope, quality, adequacy, and presuppositions, definitions or conventions regarding mathematical structures differ in scope, quality, adequacy and presuppositions.

(Depiction of CIE colour space gamuts by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image is shared under the same license.)

What is phenomenal structure, and how do we find it?

There are three important landmarks that have influenced the way in which we use mathematical structures to describe conscious experiences today: quality spaces as introduced by Austen Clark (A. Clark, 1993), quality spaces as introduced by David Rosenthal (Rosenthal, 1991, 2010) and Q-spaces as introduced in IIT 2.0 (Tononi, 2008). While these methodologies have served an important function in enabling structural research, it is also important to be clear about their shortcomings.

As far as IIT is concerned, the obvious shortcoming is that the theory does not provide a phenomenal interpretation of the structure it proposes, other than the claim that the structure "is identical to [the system's] experience" (Oizumi et al., 2014). This gives rise to what David Chalmers has called the *Rosetta Stone Problem* (Chalmers, 2023b): the problem of how to translate the mathematical structure that IIT proposes into phenomenological terms. IIT does not actually specify a methodology that clarifies how to interpret and test their proposed structure in phenomenal terms.

The proposals by Clark and Rosenthal do specify methodologies. The major shortcoming of these methodologies, on my view, is that they conflate three sources of mathematical structure:

1. **Mathematical Convenience.** Some of the structure is introduced simply for mathematical convenience.

2. **Laboratory Operations.** Some of the mathematical structure refers to, or depends on, laboratory operations.

3. **Conscious Experience.** Only part of the mathematical structure actually pertains to conscious experiences or phenomenal character.

mathematical structure physics or other natural sciences. In physics and other natural science, we do not have direct access to the phenomena we're studying. In a certain sense, for structural claims in physics, *anything goes*, as long as the relevant notion of measurement for that structure reproduces what is observed. This is why there are hugely different proposals about the structure of spacetime, for example, ranging from quantized spacetime (Rovelli, 2004) and emergent spacetime (R. Koch & Murugan, 2012) to proposals that depart completely from what we intuitively think spacetime should be (Finster & Kleiner, 2015). As long as limiting processes exist that relate these proposals to previous models, in this case the notion of spacetime of General Relativity, all those proposals are viable options. This is not the case for consciousness, because consciousness has a different epistemic context. For example, it exhibits what is sometimes called *epsitemic asymmetry*: there are "two fundamentally different methodological approaches that enable us to gather knowledge about consciousness: we can approach it from within and from without; from the first-person perspective and from the third-person perspective. Consciousness seems to distinguish itself by the privileged access that its bearer has to it" (Metzinger, 1995). In other words, in addition to the usual scientific way of accessing and modelling a phenomenon there is a second way of accessing the phenomenon (described in terms of the first person perspective *metaphor* above). Because of this different epistemic context, using mathematical structure to describe a phenomenon is different in the case of consciousness, and more constrained, than in the case of physics.

Clark's Quality Spaces

Quality spaces as introduced by Austen Clark (A. Clark, 1993) are based on the following methodology. To construct the quality space for an individual subject,[17] one fixes a class of stimuli $\S$ that can be presented to the subject, and defines two tasks that the subject can complete in response to the presentation of one or more stimuli. The first task probes whether the subject is able to *discriminate* the experience elicited by two different stimuli consciously. The second task probes whether the subject experiences a stimulus to be more similar to a reference stimulus than another stimulus. This is called *relative similarity*.[18]

The discrimination task is used to define a *global indiscriminability* relation on the class of stimuli $\S$.[19] While discriminability does not constitute an equivalence relation, global indiscriminability does. This equivalence relation partitions the set of stimuli. Each set in this partition contains stimuli which are globally indiscriminable from each other, and defines a *quality* in Clark's proposal. The collection of the sets in this partition (the space of equivalence classes of $\S$, in mathematical terms) defines the domain of the quality space that is being constructed.

The relative similarity task is used to define a graph, in the mathematical sense of the term, between the qualities: a mesh of nodes and edges that link some of the nodes. Working with stimuli that represent the different qualities, one first collects relative similarity data. This is data about whether a quality q_1 is more similar to a reference quality q_0 than another quality q_2. One might find that the pair (q_1, q_0) is more similar to each other than the pair (q_2, q_0), say. Having collected this data for all qualities in the set, one then represents them as a graph. Every quality one has previously constructed is a node of the graph, and every pair (q_i, q_j) about which one has relative similarity data is an edge of the graph between the nodes that represent the qualities. The important part now is that the edges get labels, namely numbers, and these numbers must be chosen in such a way that the relative similarity judgements that have been collected are represented truthfully by the ordering of the numbers. The label of the edge (q_1, q_0) above, for example, must be a lower number than the label of the edge (q_2, q_0), because the former pair is more similar to each other than the latter pair. The result of this procedure is a labelled graph, where the nodes represent qualities, edges indicate pairs for which similarity data is available, and labels on the edges represent relative similarity. In mathematical terms, this is called a POSET-labelled graph, where POSET means 'partially ordered set'. The partial order is the phenomenal structure of the relative similarity experiences.

[17] Clark mostly has humans in mind, but does consider the case of animals briefly in (A. Clark, 1993). Nothing hinges on humans in the methodology he proposes.

[18] There is considerable freedom in which class of stimuli to choose and how to define and implement the tasks, which is why the proposal constitutes a methodology much more than a definition, on my view.

[19] Two stimuli are globally indiscriminable if and only if the following two conditions hold:

1. The two stimuli are indiscriminable from each other.
2. The two stimuli have identical indiscriminability relations to all other stimuli in $\S$.

Up to this point all the mathematical structure is still grounded in conscious experience, to a large extent. The data to carry out the constructions is based on tasks that might utilize reports or behavioural measures, but these tasks should depend on what is experienced.

The next step in Clark's methodology consists of introducing a metric, a tool to measure distances in terms of continuous numbers, and in fact an Euclidean space that has a uniform, homogeneous metric. To this end, it makes use of a procedure known as 'multidimensional scaling' (Beals et al., 1968). In Clark's case, it consists of finding an *embedding* of the graph into an Euclidean metric space in such a way that the distance between the nodes of the graph—which are mapped to points in the metric space—reproduce the ordering of relative similarity that the labels of the graph encode.

From the perspective of phenomenal character, this step is mind-blowing. Not only is the metric introduced without any reference to experience, but this step also leads to the introduction of many more points besides the original qualities that were carefully constructed making use of global indiscriminability. Technically speaking, it leads to an infinity of additional points, all of which feature in the metric function of the space, and none of which is any different from the points that were carefully constructed based on tasks and stimuli.

The only justification I can think of why one would make use of this last step, as compared to just working with the POSET-labelled graph, is mathematical convenience. A POSET-labelled graph might just be too unfamiliar a mathematical object. Or maybe the reason is that it cannot easily be further analysed on a computer in familiar ways. These justifications are in fact made explicit in introductory texts on psychophysics. Luce and Suppes, for example, speak of representational measurement, of which multidimensional scaling is an example, as "an attempt to understand the nature of empirical observations (...) in terms of *familiar* mathematical structures" (Luce & Suppes, 2002, p. 1) (my emphasis), and add that "the use of such empirical structures in psychology is widespread because they come close to the way data are organised for subsequent statistical analysis" (Luce & Suppes, 2002, p. 2). Be that as it may, the last step that introduces the metric function fails to be grounded in conscious experience. It is an example of 1. above.

Rosenthal's Quality Spaces

The construction of quality spaces as defined by David Rosenthal is based on a class of stimuli as well. But in this case, one only needs a discrimination task, as well as means to *vary* the stimuli.

The main step in Rosenthal's methodology is to construct *Just Noticeable Differences* (JNDs) from variations of the stimuli and the discrimination task. To this end, one varies a stimulus in some direction until the subject notices the difference between the stimulus and the variation. The class of stimuli which one can reach by varying one stimulus without creating a JND gives a set or region in stimulus space, and much like in the case of Clark, the idea is that these regions constitute qualities. A metric function is

introduced on the set of qualities by counting the minimal number of regions one has to pass so as to go from one quality to the other.

In this proposal too, there is a question as to the experiential source of the metric function. Because the metric function can be specified once JNDs have been constructed without needing any additional data, it might not legitimately represent anything over and above the JNDs and their neighbourhood relations. Furthermore, while we do experience color qualities as instantiating a relative similarity structure, we do not experience qualities to be a certain number of steps apart, as a metric would require if it indeed represented a structure of conscious experience.[20] So there is a worry of the metric being due to mathematical convenience as well here.

A more fundamental worry though in this case concerns the *variations* of stimuli that one needs in order to construct JNDs and their neighborhood relations in the first place. The idea of a variation—starting with one stimulus and then changing that stimulus continuously until a subject notices a difference—requires a *topology* on the stimulus space. A topology defines what it means to "draw a line without lifting a pen" on an abstract space, so to speak. And it is precisely what provides the continuous curves required for variations. Without a topology, there is no notion of closeness of two points. One could go from any point to any other point immediately.

The problem is that different topologies give different variations. So when one actually constructs a quality space according to Rosenthal's methodology in the lab, the resulting space depends on the topology of the stimulus space that has been used. Much like there isn't just a single notion of colour space, there isn't just a single topology on colour stimuli one can use. As a result, the metric function that one constructs in an application of Rosenthal's methodology actually depends on the topology that has been chosen in the experiment, which is a laboratory operation in the sense of 2. above.

In the case of Rosenthal's methodology, there is in fact a theory that can be used to answer these and similar worries, a theory about what consciousness is, about how qualities should be understood, and about how consciousness and qualities relate. When I asked David Rosenthal about the problem regarding topology, for example, he countered by assuming that there is just one actual physical topology in reality and that this is the topology that should be used. It is not clear to me how this would work in practice, given that this topology is presumably defined by Quantum Electrodynamics (QED), and too far removed from experimental practice to be applicable—in the lab, some choice of topology will have to be made nonetheless—, but theoretically speaking, the answer is fully valid. Similarly, the theory about what qualities are and how they relate to consciousness discharges the methodology from the problem that, according to the subsumed notion of discrimination in this case, discriminations could also be made unconsciously.

There is, however, no free lunch. And the price to be paid for solving methodological problems by theoretical assumptions is that the methodology now depends on these assumptions and cannot be used to formulate or test other theories of consciousness.

[20] For a more careful examination of the case of a metric, cf. (Kleiner & Ludwig, 2024). For questions on how quality spaces *should* relate to consciousness or phenomenal character according to the underlying theory, cf. below.

The methodological tool might be deprived of much of the impact it could otherwise have.

On my view, quality spaces are ways to describe or represent the explanandum—what is to be explained: qualitative or phenomenal character, what it is like to be—, while theories of consciousness are the explanantia—what does the explaining. This is why I have always been tempted to read Rosenthal's proposal as a general methodology that is independent from his theory. This is possible and addressing the above-mentioned problems on purely methodological grounds leads, on my view, to fruitful further developments of his construction (cf. below and (Kleiner & Ludwig, 2024)).

How to move forward

In the last two sections, I have analysed two proposals for methodologies that define what quality spaces are. While these proposals have served an important role in enabling structural thinking, much of the essential structure in these proposals is not actually grounded in conscious experiences, but in mathematical convenience and laboratory operations.

It is possible to go beyond individual methodologies and analyse the *type of condition* that is applied in these proposals and more recent work. That is, the type of condition that decides whether a mathematical structure is a quality space or phenomenal space—*a mathematical structure of conscious experience*, to use a general term. In a nutshell, all existing proposals I know of amount to:

(A) Conditions on the domains (sets of points) of a mathematical structure, formulated in terms of qualities, qualia, phenomenal properties or similar aspects of conscious experiences.

(B) The requirement that the mathematical axioms of the structure (such as the axiom that the metric distance between a point and itself is zero) are satisfied.

This type of condition can be shown to be insufficient to ground a thorough understanding of phenomenal structure. This is the case because (a) it is prone to admitting incompatible structures, (b) allows for *arbitrary* re-definitions of structures that still satisfy the condition, and (c) in a subtle but important sense, the condition is indifferent to structural facts of conscious experience. I do not have the space here to explain these problems in detail; they are explained and illustrated in (Kleiner & Ludwig, 2024, Section 1).

I take the problems of existing proposals, and the insufficiency of the general type of condition that is applied, to constitute a need of constructing a *new methodology* for phenomenal spaces. This methodology needs to take previous methodologies into account, but needs to amend and extend them to avoid the three insufficiency problems as well as the issues with non-conscious sources of the mathematical structure.

In (Kleiner & Ludwig, 2024), Tim Ludwig and I have set out to find a methodology that achieves this task. The result is illustrated in Figure 7.5.2. The proposal shares

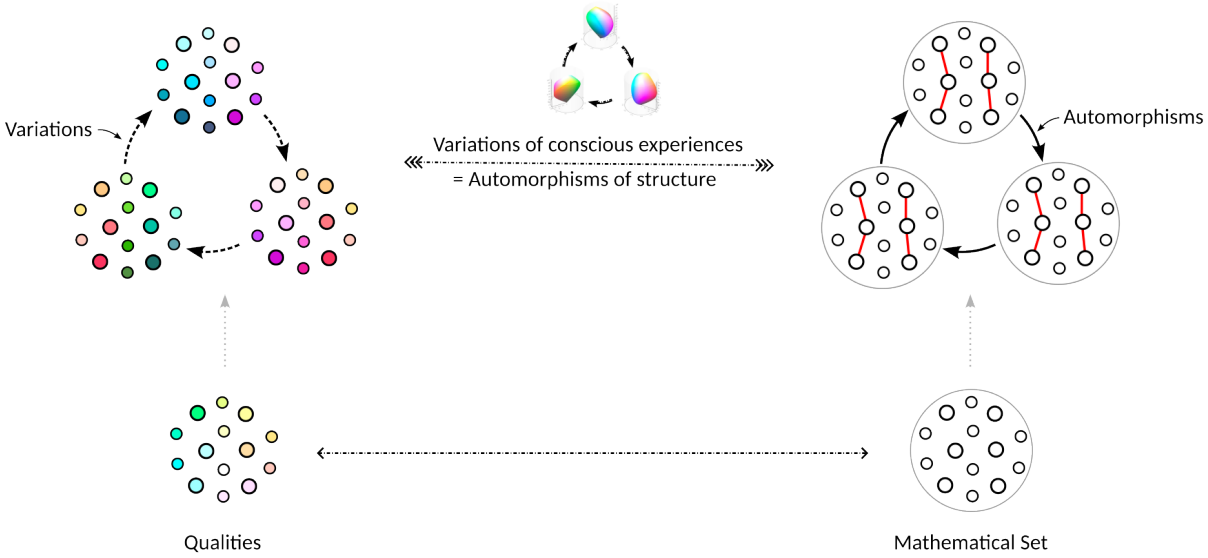


Figure 7.5.2.: **How to define phenomenal spaces?** This figure illustrates one way to define phenomenal spaces and other mathematical structures of conscious experiences. One starts out with a choice of qualities (bottom left), for example colour qualities, sometimes also called qualia or conceptualised as instantiated phenomenal properties. The qualities form a set that constitutes the points of the phenomenal space (bottom right). Every experience comprises a subset of qualities, and as experiences change from one experience to the next, the subset of qualities that is realised varies (top left). These variations can be understood as mappings from the set of qualities to itself, and therefore have the same formal structure as automorphisms (Fig. 7.2.1): mappings from the points of a space to other points of the space (top right). This allows for the following simple definition of phenomenal structure: Phenomenal structure is that mathematical structure whose automorphisms are identical to the variations of the qualities as experiences change. Put differently, phenomenal structure (indicated here by red lines) is that mathematical structure which renders the following statement true: the variations of (qualities of) conscious experiences are the automorphisms of the structure (top centre). For details, cf. (Kleiner & Ludwig, 2024).

with David Rosenthal's methodology that it rests on variations, though in our case, any transition from one conscious experience to another counts as variation, and we do not demand continuity or restrict only to variations of stimuli.

Put in terms of phenomenal properties, the core intuition of our proposal is that a mathematical structure is a phenomenal space if and only if there is a phenomenal property that behaves exactly as the mathematical structure does under variations. If a variation preserves the mathematical structure (if it is an automorphism of the structure, in mathematical terms), then it must not change the phenomenal property. If, conversely, a variation does not preserve the mathematical structure, then it must change the phenomenal property. In a nutshell: there is something "in" conscious experience (the phenomenal property) that behaves exactly as the mathematical structure does.

7.6. Conclusion

Structural approaches, which make use of mathematical structure to describe or model conscious experiences, offer new and valuable avenues for studying consciousness. My aim in this paper is to provide three comments that I consider important when engaging in structural research. Each comment targets what is, in my view, a misconception or misunderstanding that I aim to clarify.

My first comment focuses on the metaphysical underpinnings of structural approaches. I show that, contrary to popular belief, structural approaches are not tied to physicalist or reductive metaphysics. Instead, they offer versatile descriptive tools that can be utilised irrespective of one's metaphysical commitments, across research programmes of any metaphysical flavour.

My second comment concerns isomorphisms and structure-preserving mappings. A number of emerging structuralist research programmes rely on assuming a structure-preserving mapping between the phenomenal and the physical domain. I argue that this assumption is unwarranted, and that isomorphisms and structure-preserving mappings are not the right mathematical object to provide explanations, predictions, or definitions of phenomenal structure. Instead, we should direct our attention to structural theories of consciousness, without expecting a single mathematical formalism to fit them all. One major experimental consequence of this is that methods such as Representational Similarity Analysis (Kriegeskorte, Mur, & Bandettini, 2008), which search for structural similarity, may not be the right approach to search for the neural correlates of structure.

My third and final comment focuses on the question of what phenomenal structure *is*, and how we find it. Conscious experiences do not "come with" mathematical structure in any meaningful sense. Rather, mathematical spaces and mathematical structure offer a language to describe or represent conscious experiences, and just like we need definitions or conventions to apply English language terms to consciousness, we need definitions or conventions to apply structural terms. In the case of structure, the definitions and conventions take the form of methodologies that govern how to construct or use the mathematical terminology. In my final comment, I review the two major method-

ologies that have guided recent developments: quality spaces as introduced by Austen Clark, and quality spaces as introduced by David Rosenthal. I show that both suffer from fundamental issues, and discuss how to move forward in light of this.

8. What is a Mathematical Structure of Conscious Experience?

Johannes Kleiner, Tim Ludwig[1]

Attempts to represent conscious experiences mathematically go back at least to 1860 (Fechner, 1860), and a large number of approaches have been developed since. They span psychophysics, philosophy, phenomenology, neuroscience, theories of consciousness, and mathematical consciousness science (A. Clark, 1993; Stanley, 1999; A. Clark, 2000; Yoshimi, 2007; R. G. Kuehni & Schwarz, 2008; Rosenthal, 2010; Klincewicz, 2011; Kostic, 2012; Zaidi et al., 2013; Mason, 2013; Hoffman & Prakash, 2014; Oizumi et al., 2014; Renero, 2014; Young et al., 2014; Rosenthal, 2015, 2016; Gert, 2017; Grindrod, 2018; A. Haun & Tononi, 2019; Prentner, 2019; Kleiner, 2020b; Fortier-Davy & Millière, 2020; Lee, 2021; Tsuchiya & Saigo, 2021; Coninx, 2022; Lee, 2022; Resende, 2022; Tallon-Baudry, 2022; Tsuchiya et al., 2022) and are known under various different names, including quality spaces (A. Clark, 1993), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021), Q-spaces (Chalmers & McQueen, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink et al., 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). The mathematical structures and spaces introduced by these approaches have enabled significant advancements in their respective fields. Nevertheless, this research remains largely fragmented. The various approaches employ different formalizations and different mathematical structures, and they presume a different, and sometimes partial, understanding of the concept of a mathematical structure or space when applied to conscious experience. What is missing, from our perspective, is a definition of the term 'mathematical

[1]Published as: Kleiner, J., & Ludwig, T. (2024). What is a mathematical structure of conscious experience?. *Synthese*, 203(3), 89. (Kleiner & Ludwig, 2024)

structure of conscious experience' that clarifies how this term can and should be used.

In this article, we propose a definition of mathematical structures of conscious experience. Our main desideratum is that for a mathematical structure to be *of* conscious experience, there must be something *in* conscious experience that corresponds to that structure: a specific structural aspect of conscious experience.

Our key idea is to use variations to identify and investigate these structural aspects of conscious experience. That is because variations can serve as a binding link between conscious experiences and mathematical structures: on the one hand, variations relate to conscious experiences, because variations change aspects of conscious experiences (like qualia, qualities, or phenomenal properties); on the other hand, variations relate to mathematical structures, because they may or may not preserve them.

In defining a mathematical structure of conscious experience, our proposal does not answer the question of what this mathematical structure actually is, or which type it has. Instead, our proposal identifies the analysandum for future work on spaces and structures of conscious experience, based on which phenomenal spaces, quality spaces, qualia spaces, Φ-structures, as well as several other related concepts, can be constructed and investigated.

This paper is structured as follows. In Section 8.1, we discuss how recent approaches relate mathematical structures to conscious experience and identify three key issues in these approaches. In Section 8.2, we present our proposal together with the necessary background information. In Sections 8.3, and 8.4, we consider two important examples: relative similarity and topological spaces. In Section 8.5, we show how our proposal resolves the three problems identified in Section 8.1. Finally, our conclusion follows in Section 8.6.

8.1. The Status Quo

So where do things stand? Most of the early work that has attributed mathematical structure to conscious experience was grounded in intuition. Whether or not a specific mathematical structure is a mathematical structure of conscious experience–a structure which "pertains to", or "belongs to" consciousness, that is–was not assessed systematically; instead, it was assessed based on an intuitive insight of appropriateness. More recent approaches have realized the need for a more systematic method, for example (Gert, 2017; Lee, 2021, 2023; Prentner, 2019; Resende, 2022; Rosenthal, 2015, 2016). In this section, we analyze what we take to be the condition that underlies these approaches: a condition that justifies prescribing a mathematical structure to conscious experience. As we will see, this condition is quite natural. But, as we will demonstrate, it cannot be understood as a sufficient condition.

In a nutshell, a mathematical structure consists of two building blocks; for a detailed introduction, see Section 8.2.2. The first building block consists of one or more sets called the *domains* of the structure. The second building block are *relations or functions* which are defined on the domains. For reasons explained below, we will denote

them as *structures* in the narrow sense of the term. A metric space, for example, is a mathematical structure that is defined on the two domains: a set of points and the real numbers. Furthermore, it comprises a function—the so-called metric function—which maps two points to a real number. A topological space, to give another example, is a mathematical structure that is defined on a single domain: a set of points. Furthermore, it comprises a collection of unary relations, which are subsets of the domain.[2]

Usually, a mathematical structure also comes with *axioms*. The axioms establish conditions that the functions or relations have to satisfy. In the case of a metric structure, the axioms require the metric function to satisfy three conditions, called positive definiteness, symmetry, and triangle inequality. In the case of a topological structure, the axioms ensure the collection includes the empty set and the whole domain, that it is closed under finite intersections, and that it is closed under arbitrary unions.

When put in these terms, recent proposals that go beyond intuitive assessments, make use, either directly or indirectly, of the following condition to justify that a specific mathematical structure is a mathematical structure of consciousness. Here, we use the term *aspect* as a placeholder for qualia, qualities, (instantiated) phenomenal properties, or similar concepts.[3]

(MDC) A mathematical structure is a mathematical structure of conscious experience if and only if the following two conditions are satisfied:

(D1) The domains of the structure are sets whose elements correspond to aspects of conscious experiences.

(D2) The axioms of the structure are satisfied.

In the case of the metric structure introduced in (A. Clark, 1993), for example, (D1) is satisfied because the set of points corresponds to qualities of conscious experience. The real numbers might have a phenomenal interpretation as describing degrees of similarity, as for example in (Lee, 2021). Condition (D2) requires positive definiteness, symmetry, and the triangle inequality to hold. This includes, for example, the condition that "points should have distance zero just in case the qualities represented by those points are phenomenally identical" (Lee, 2021, p. 14). In the case of the topological structure introduced in (Stanley, 1999), to give another example, (D1) is satisfied because the domain of the structure refers to qualia. Condition (D2) would require, then, that the chosen collection of subsets satisfies the axioms of a topological space.

Prima facie, (MDC) could be taken to define what a mathematical structure of conscious experience is. However, if understood as sufficient condition, the following three

[2] A unary relation on a domain, in the mathematical sense, is a subset of the domain; see Section 8.4.

[3] We use the term 'aspect' as a placeholder for these terms because the above condition is not unanimously framed in either of these terms, and because our proposal in Section 8.2 is applicable with respect to any of these choices. In short, our goal is not to pick any one of these concepts but to offer a definition that works with respect to any of these concepts. Which concept is best suited for a particular task or domain is a philosophical question that can be answered independently of our proposal.

problems arise.

Problem 1: Incompatible Structures

A first reason why (MDC) cannot be a sufficient condition to assess whether a mathematical structure is a mathematical structure of consciousness is that it allows for incompatible structures.

Consider, as an example, the case of topology. A basic question in topology is whether a target domain is discrete or not. A target domain is discrete if and only if its topology contains all subsets of the domain (K. Joshi, 1983). Otherwise, the target domain is not discrete. These two cases are exclusive, meaning that discrete and non-discrete topological structures are incompatible.

According to (MDC), conscious experience has a discrete structure. That is because any set whatsoever can be equipped with the discrete topology. Therefore, picking a set X of aspects (qualia, qualities, phenomenal properties, etc.) and choosing its discrete topology provides a mathematical structure that satisfies both conditions (D1) and (D2). But, according to (MDC), consciousness also has a non-discrete structure. That is because any set can also be equipped with a non-discrete topology. We can, for example, take an arbitrary decomposition of the set X into two subsets A and $A^{\perp}$, where $A^{\perp}$ is the complement of A, and consider the topology $\{\emptyset, A, A^{\perp}, X\}$. This choice satisfies all axioms of a topology, and therefore satisfies (D2). Furthermore, it is built on the same set X as the discrete topology above, which implies that it also satisfies (D1). Therefore, the discrete and the non-discrete topological structures are both structures of conscious experience, according to (MDC).

This example shows that, if understood as a sufficient condition, (MDC) implies that two incompatible structures are both structures of conscious experience, and that they do so with respect to the exact same domain of aspects. The condition fails to determine which of the two incompatible structures is the right one.

Problem 2: Arbitrary Re-Definitions.

A second reason why (MDC) cannot be a sufficient condition is that it allows for arbitrary re-definitions: if one structure is given that satisfies (MDC), then any arbitrary definition of a new structure in terms of the given structure also satisfies (MDC), so long as the domains of the structure remain unchanged. If the former pertains to consciousness, so does the latter.

A simple and well-behaved example of this is given by rescaling a metric function. Let us suppose that (M, d) is a metric structure which pertains to consciousness according to (MDC), where M is a set of aspects and d is the metric function, which provides a real number $d(a, b)$ for every two aspects a and b. Since (M, d) satisfies (MDC), so does every structure $(M, C \cdot d)$, where $C \cdot d$ is the multiplication of the function d by a positive real number C. Here, the number C can be chosen arbitrarily. If one metric structure

pertains to consciousness according (MDC), so does an uncountably infinite number of metric structures.

What is more, when re-defining structures, one is free to change the axioms as one pleases. For example, we could pick any function f that maps M to the positive real numbers and define a new distance function by $(f(a) + f(b)) \cdot d(a, b)$. This is not a metric structure anymore, because the triangle inequality axiom does not hold. But it still satisfies positive definiteness and symmetry, and therefore satisfies (MDC), with a new set of axioms. One could even break asymmetry to get a distance function like the one applied by IIT (Kleiner & Tull, 2021). More severe cases appear with more complicated structures.

This is a problem, not only because of the unlimited number of structures that appear, but also because there is an arbitrariness in the definition of a new structure, specifically concerning the axioms. It seems strange that the axioms can be redefined at will, so as to always satisfy Condition (D2). Something is missing that restricts this arbitrariness in (MDC).

Problem 3: Indifference to Consciousness.

The third reason, which speaks against the sufficiency of (MDC), is that the proposed condition seems somewhat indifferent to details of conscious experience.

To illustrate this indifference, let us consider again the discrete and non-discrete topological structures from above. As we have shown, these structures pertain to conscious experience according to (MDC). Yet, nothing more than a few lines needed to be said to establish this fact. In particular, we did not need to use any noteworthy input related to consciousness other than picking some set of aspects; and it didn't matter which aspects we picked.

It is a red flag if so short an analysis, which does not depend on consciousness in a meaningful way, establishes facts about the mathematical structure of conscious experience. The example exposes an indifference of (MDC) to details of conscious experience: the definition only relates to the different aspects, but not to the sort of mathematical object that connects these different aspects. Speaking somewhat vaguely, (MDC) does not refer to the "way" in which the different aspects of consciousness are related. This is why, in the case of topology, it allows one to draw conclusion without any noteworthy input from actual experience. This constitutes another reason that condition (MDC) is missing some important component, if used as sufficient condition.

Cause of these Problems

These three problems arise because (MDC) is not only a necessary, but also a sufficient condition: it contains an 'if' condition in addition to the 'only if' condition. In the first example, we show that two incompatible mathematical structures—a discrete and a non-discrete topology—each satisfy (D1) and (D2). Because (MDC) is a sufficient condition, it follows that both structures are structures of conscious experience, according

to (MDC). In light of the incompatibility of discrete and non-discrete topologies, this constitutes an issue of the definition. In the second problem, we show that for any given structure or space that satisfies (D1) and (D2), any arbitrary redefinition yields a structure or space which also satisfies (D1) and (D2), for a suitably adapted set of axioms. Because (MDC) is a sufficient condition, this implies that the arbitrarily redefined structure is also a mathematical structure of conscious experience, which for reasons explained above, constitutes an issue as well. The third example, finally, builds on the first example and makes use of the sufficient condition in exactly the same way. Because there is no condition in (MDC) that relates to structure in the narrow sense of the term—no condition that relates to relations or functions, that is—, and because of the sufficient condition in (MDC), structures of conscious experience can be established without reference to structure in the narrow sense of the term.

The Way Forward

To resolve the three problems, our task is to propose a definition for a mathematical structure of conscious experience that makes sense as a necessary and sufficient condition. This will be the content of Section 8.2.

Two desiderata guide our search. First, as is the case with (MDC), an improved definition should be *about* conscious experience in the sense that it targets qualities, qualia, instantiated phenomenal properties, or similar aspects of conscious experience, as in (D1) above. Second, there should be something *in* conscious experience—a quality, or quale or phenomenal property—that relates to structure in the narrow sense of the term. This "something" should make sure that the definition is not indifferent to conscious experience in the sense of Problem 3, and that the definition refers to functions or relations in a meaningful way, so as to stop arbitrary re-definitions (Problem 2). The proposal which we present in the next section is the result of our search.

Despite the above-mentioned problems, we think that (MDC) is an important condition. It might not be suitable as a sufficient condition, but it is valuable as a necessary condition. If one understands mathematics pragmatically as constituting a *language*– a body of symbols and terms with rules that connect these–, then mathematics can be used to describe phenomena, much like the English language can. Looking back at Condition (MDC) after our analysis, and presuming this pragmatic conception of mathematics, we think that (MDC) is best understood as an expression of what it takes for a mathematical structure to *describe* conscious experience. That is, (MDC) might be a valuable descriptive tool that utilizes mathematical structure to represent information on how aspects are related to each other (as explicated by (D1) and (D2)).

Because of this, we will refer to a mathematical structure that satisfies (MDC) as a mathematical structure that 'describes conscious experience' in what follows. The new condition which we develop below contains (MDC) as necessary part; this is aligned with the intuition that any mathematical structure of conscious experience also describes conscious experience.

8.2. Mathematical Structures of Conscious Experience

In this section, we provide a definition of what mathematical structures of conscious experience are. Based on this definition, phenomenal spaces, quality spaces, qualia spaces, and related structures can be constructed and investigated. The definition embodies a way to think and work with mathematical structures when applied to conscious experience.

Our key desideratum in improving (MDC), explained above, is that for a mathematical structure to be a mathematical structure of conscious experience, rather than just a descriptive tool for conscious experience, there must be a structural aspect in conscious experience that relates to that structure. A major goal of this section is to explain this in detail. Denoting a mathematical structure by S, we call this structural aspect an S-aspect.

To make sense of what an S-aspect is, we need to understand how aspects (like qualia, qualities or phenomenal properties) relate to mathematical structures. While aspects may have an arity, meaning they may be instantiated relative to other aspects, they are not experienced as having a mathematical structure per se (unless, of course, they are aspects of experiences of mathematical structures themselves, such as of geometric shapes). Therefore, relating aspects to mathematical structures requires a tool that applies both; concrete aspects of conscious experience and abstract formal entities. Variations provide such a tool.

In general, a variation is a change of something into something else; in our case, it is a change of one experience into another experience. Such variations may be induced by external stimuli or interventions, occur naturally, or be subjected to imagination ('imaginary variations' (Husserl, 1936)). Variations are directly related to aspects of conscious experiences because a variation can change an aspect. This is the case iff an aspect is part of the experience before the variation but isn't part of the experience after the variation. And variations are also intimately related to mathematical structures, because they may or may not preserve them, as explained in detail below. An S-aspect, then, is an aspect that is changed by a variation if and only if the variation does not preserve the structure S. To explain this in detail is the purpose of the remainder of this section.

8.2.1. Terminology and Notation

Here, we introduce the key terms we use to define mathematical structures of conscious experience. These terms are *conscious experiences*, *aspects* of conscious experiences, and *variations* of conscious experiences. The introduction proceeds axiomatically, so that our construction does not rely on a specific choice of these concepts. Rather, any choice of these concepts that is compatible with what we say here can be the basis of an application of our definition.

Our construction is based on a *set E of conscious experiences* of an experiencing subject. We denote individual conscious experiences in that set by symbols like e and e'; formally $e, e' \in E$. From a theoretical or philosophical perspective, one may think of

the set E as comprising all conscious experiences which one experiencing subject can have, i.e. all nomologically possible experiences of that subject. From an experimental or phenomenological perspective, one may think of this set as comprising all conscious experiences that can be induced in the lab or in introspection. Different such choices may lead to different mathematical structures being accessible.

We use the term *aspect* as a placeholder for concepts such as *qualia* (Tye, 2021), *qualities* (A. Clark, 1993), *mental qualities* (Rosenthal, 2010), or (instantiated) *phenomenal properties*.[4] For every experience $e \in E$, we denote the set of aspects instantiated in this experience by $A(e)$. The set of all aspects of the experiences in E, denoted by $\mathcal{A}$, is the union of all $A(e)$; formally $\mathcal{A} = \bigcup_{e \in E} A(e)$. Individual aspects, that is members of $\mathcal{A}$, will be denoted by small letters such as a, b, c. When explaining examples, we will often use the abbreviation 'a is the experience of ...' as a shorthand for saying 'a is a ... aspect of an experience'. For example, 'a is the experience of red color' means 'a is a red color aspect of an experience'.

Some aspects may require other aspects for their instantiation. For example, it is usually the case that an experience of relative similarity is an experience of relative similarity of something, for example two color aspects relative to a third color aspect. If an aspect a requires other aspects for its instantiation, we will say that the aspect a *is instantiated relative to* aspects $b_1, ..., b_m$, or simply that a *is relative to* $b_1, ..., b_m$. Aspects which are instantiated relative to other aspects are the building blocks for the structure of conscious experience.

A *variation* of a conscious experience e changes e into another experience e'. Because experiences have structure, there may be various different ways to go from e to e'.[5] Therefore, in addition to specifying e and e', a variation is a partial mapping

$$v : A(e) \to A(e') \, .$$

This mapping describes how aspects are replaced or reshuffled by the variation. A mapping which is not surjective, meaning that it does not map to all aspects in $A(e')$, makes room for appearance of new aspects. A mapping which is partial, meaning that it does not specify a target for every aspect in $A(e)$, makes room for aspects to disappear.

[4] Many other concepts work as well. For example, if one works with an atomistic conception of states of consciousness, where the total phenomenal state of a subject—what it is like to be that subject at a particular time—is built up from individual atomic states of consciousness, one can take e to denote the total phenomenal state and aspects to be the *states of consciousness* in that total state. Another example would be to take aspects to denote phenomenal distinctions as used in Integrated Information Theory (Tononi, 2015). What matters for our definition to be applicable is only that according to one's chosen concept of conscious experience, every conscious experience exhibits a set of aspects.

[5] To illustrate this point, consider, for example, the following two mappings v and v' which map the numbers 1, 2, and 3 to the numbers 2, 4, and 6. The mapping v is the multiplication of every number by 2, meaning that we have $v(1) = 2$, $v(2) = 4$, $v(3) = 6$. The mapping v', on the other hand, is defined by $v(1) = 6$, $v(2) = 2$, $v(3) = 4$. If we only cared about the *sets* of elements that these mappings connect, the mappings would be equivalent: there is no difference between the set $\{2, 4, 6\}$, which is the image of v, and $\{6, 2, 4\}$, which is the image of v'. If, however, we care about the *structure* of the elements of the sets–in this case, the *ordering* of numbers–, then there is a difference. While $2 \leq 4 \leq 6$, it is not the case that $6 \leq 2 \leq 4$. Because we care about the order of the elements, we need to say which element goes where.

8.2.2. What is a Mathematical Structure?

To find a rigorous definition of the mathematical structure of conscious experience, we need to work with a rigorous definition of mathematical structure. Mathematical logic provides this definition, which we now review.

A *mathematical structure* $\mathbb{S}$ consists of two things: domains, on the one hand, and functions or relations, on the other hand. We now introduce these concepts based on two simple examples.

The *domains* of a structure $\mathbb{S}$ are the sets on which the structure is built. We denote them by $\mathcal{A}_i$, where i is some index in a parameter range I. In the case of a metric structure, for example, the domains would be $\mathcal{A}_1 = M$ and $\mathcal{A}_2 = \mathbb{R}$, where M is a set of points and $\mathbb{R}$ denotes the real numbers, understood as a set. In the case of a strict partial order, there is just one domain $\mathcal{A}$, which contains the elements that are to be ordered.

The second ingredient are functions and/or relations. *Functions* f map some of the domains to other domains. In the case of a metric structure, the function would be a metric function $d : M \times M \to \mathbb{R}$, which maps from $\mathcal{A}_1 \times \mathcal{A}_1$ to $\mathcal{A}_2$. A *relation* R, in the mathematical sense, is a subset of the m-fold product $\mathcal{A}_i \times ... \times \mathcal{A}_i$. Here, $\mathcal{A}_i$ is the domain on which the relation is defined, and m is the arity of the relation, which expresses how many relata the relation relates. The product is usually just written as $\mathcal{A}_i^m$. In the case of a strict partial order, the relation is binary, which means that R is a subset of $\mathcal{A}^2$. For binary relations, one usually uses notation like $a < b$ instead of writing $(a, b) \in R$.

In almost all cases, mathematical structures also come with *axioms*, which establish conditions that the functions or relations have to satisfy. They are useful because they constrain and classify the structure at hand. For $\mathbb{S}$ to be a metric structure, for example, the function d has to satisfy the axioms of positive definiteness, symmetry, and triangle inequality (Rudin, 1976). For $\mathbb{S}$ to be a strict partial order, the relation R has to be irrefelxive, asymmetric, and transitive (K. D. Joshi, 1989).

To have a nice and compact notation, we will use one symbol S_j to denote both functions and relations. That is because, in any concrete proposal, it is always clear whether S_j is a function or a relation.[6] The index j takes values in some parameter range J that specifies how many functions or relations there are. Using this notation, we can represent the definition of mathematical structure provided by mathematical logic as follows:

A mathematical structure $\mathbb{S}$ is a tuple

$$\mathbb{S} = \big((\mathcal{A}_i)_{i \in I}, (S_j)_{j \in J}\big)$$

of domains $\mathcal{A}_i$ and functions or relations S_j.

For given domains $\mathcal{A}_i$, the mathematical structure $\mathbb{S}$ is fully determined by the S_j. Thus, we can also refer to S_j as 'structures', if the domains are clear from context. For

[6] In mathematical logic, mathematical structures are denoted as triples of domains, relations, and functions. However, in our case, using just one symbol for functions and relations improves readability substantially.

simplicity, we can drop the index j and simply write S whenever we consider just one such structure.

As a final step in this section, we introduce the *relata* of a structure S. This will be helpful to write things concisely below. The term relata designates those elements that are related by a structure. In the case where S is a relation R on a domain $\mathcal{A}$ and has arity m, these are the elements of the m-tuples $(b_1, ..., b_m) \in R$. In the case where S is a function $f : \mathcal{A}_1 \times ... \times \mathcal{A}_{m-1} \to \mathcal{A}_m$, the relata are the elements of the m-tuples $(b_1, ..., b_{m-1}, b_m)$ where $b_m = f(b_1, ..., b_{m-1})$, and where the other b_i range over their whole domains. For notational simplicity, we write $b_1, ..., b_m$ instead of $(b_1, ..., b_m)$ when designating relata below.

8.2.3. What is a Mathematical Structure of Conscious Experience?

Finally, to the heart of the matter! We recall that we have so far identified two desiderata for a mathematical structure $\mathbb{S}$ to be a mathematical structure of conscious experience. First, it should be *about* conscious experiences in the sense that its domains should correspond to aspects of conscious experiences. Second, there should be aspects *in* conscious experience that relate to the structure $\mathbb{S}$. The following definition satisfies these two desiderata. Its explanation is the task of the remainder of this section.

(MSC) A mathematical structure $\mathbb{S}$ is a mathematical structure of conscious experience if and only if the following two conditions hold:

(S1) The domains $\mathcal{A}_i$ of $\mathbb{S}$ are subsets of $\mathcal{A}$.

(S2) For every S_j, there is a S_j-aspect in $\mathcal{A}$.

Here, $\mathcal{A}$ denotes the set of all aspects of the experiences in E; formally $\mathcal{A} = \bigcup_{e \in E} A(e)$, the $\mathcal{A}_i$ denote the domains of the structure $\mathbb{S}$, and the S_j-aspects are defined below.

Condition (S1) guarantees that the first desideratum is satisfied. Condition (S2) guarantees that the second desideratum is satisfied. Furthermore, whenever a certain *type* of structure (metric, topological, partial order, manifold, etc.) is claimed to be a structure of conscious experience, the axioms that constrain and classify that type have to hold. Therefore, any mathematical structure of conscious experience (MSC) is also a mathematical structure that describes conscious experience according to (MDC). The condition that has been applied in previous proposals remains a necessary condition in (MSC).

The remaining task of this section, then, is to explain what an S_j-aspect is. For notational simplicity, we use the symbol S to denote S_j. As we have emphasized before, variations are key to understand the structure of conscious experience, because they link aspects and structure. Therefore, to be able to precisely define what an S-aspect is, we need to understand how variations relate to aspects, on the one hand, and structures, on the other hand. Our strategy is to first discuss how variations relate to aspects.

This amounts to specifying what precisely it means for a variation to change an aspect. Second, we focus on how variations relate to mathematical structure. This amounts to explaining what it means for a variation to preserve a structure. Finally, combing these two steps allows us to understand S-aspects and provide a useful definition.

What does it mean for a variation $v : A(e) \to A(e')$ to change aspects? The underlying idea is simply that an aspect is present in the source of the variation, $A(e)$, but not present any more in the target of the variation, $A(e')$. We need to take into account, though, that aspects are often instantiated relative to other aspects (see Section 8.2.1). This can be done as follows.

A variation $v : A(e) \to A(e')$ **changes an aspect** $a \in A(e)$ *relative to* $b_1, ..., b_m \in A(e)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(e)$, but a is not instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$.

In the case where $a \in A(e)$ is not instantiated relative to other aspects, the definition indeed reduces to the simple condition that $a \in A(e)$ but $a \notin A(e')$. The negation of the definition is also as intuitively expected: the aspect is present both in the source and in the target.[7]

For applications it is important to understand that this definition can fail to apply in two ways. First, it can fail because there is no a in $A(e')$ which is instantiated relative to $v(b_1), ..., v(b_m)$. This, in turn, can be the case either because there is no a in $A(e')$ at all, or because there is an a in $A(e')$ but it is instantiated relative to other aspects. Second, it can fail because one or more of the $v(b_1), ..., v(b_m)$ do not exist. The second case is possible because v is a *partial* mapping, which means aspects can disappear.

What does it mean for a variation to preserve a mathematical structure? The underlying idea is that a variation preserves the structure if and only if the structure is satisfied before the variation and remains to be satisfied after the variation. By its very nature, this is a mathematical condition, namely the condition of being a homomorphism (Mileti, 2022). The definition of a homomorphism, though, always applies to all elements of a domain at once. For our case, it is best to refine this definition to a single set of relata.[8]

A variation $v : A(e) \to A(e')$ **preserves a structure** S *with respect to relata* $b_1, ..., b_m \in A(e)$ if and only if we have

(P1) $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ if S is a relation R, or

(P2) $v\big(f(b_1, ..., b_{m-1})\big) = f\big(v(b_1), ..., v(b_{m-1})\big)$ if S is a function f.

[7] Because the definiendum already includes the first part of the condition, the negation is as follows:
A variation $v : A(e) \to A(e')$ *does not change an aspect* $a \in A(e)$ *relative to* $b_1, ..., b_m \in A(e)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(e)$ and a is also instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$.
We felt that is the best way of writing things to optimize clarity.

[8] For notational simplicity, we write $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ instead of $R\big(b_1, ..., b_m\big) \Leftrightarrow R\big(v(b_1), ..., v(b_m)\big)$.

As in the previous case, the negation of this definition is exactly what is intuitively expected: a variation does not preserve the structure if and only if the structure is satisfied before the variation, but not satisfied after the variation.[9]

For applications it is again important to see that the definition can fail for two reasons. First, it could be the case that one or more of the $v(b_i)$ do not exist in $A(e')$, if the corresponding aspect disappears. Second, the identities may fail to hold.

We now have the keys to understand S-aspects. The underlying idea is that an S-aspect is an aspect that, under any variation, behaves exactly as the structure S does: whenever S is preserved, the S-aspect does not change, and whenever the S-aspect changes, the structure S is not preserved. This is expressed by the following definition.

An aspect $a \in \mathcal{A}$ is an S-aspect if and only if the following condition holds:
A variation *does not preserve* S with respect to relata $b_1, ..., b_m$ *if and only if* the variation *changes* a relative to $b_1, ..., b_m$.

Here, the condition needs to hold true for all variations and all relata. This means that it needs to hold true for all variations of all experiences e in the set E that instantiate relata of the structure S.

This concludes our proposal for the definition of the mathematical structure of conscious experience. It is a structure whose domains correspond to sets of aspects, and which contains an S-aspect for every relation or function of the structure. In the next two sections, we apply this definition to examples. On the one hand, these examples illustrate the definition. On the other hand, they provide new insights to structures that have been featured prominently in previous approaches.

8.3. Relative Similarity

Our first example concerns relative similarity, which plays an important role, for example, in the construction of quality spaces by Austen Clark (A. Clark, 1993, 2000).

A first step in applying our definition is to choose a set E. Here we take E to comprise experiences of three color chips, as indicated in Figure 8.3.1A, where one of the chip (the reference) has a fixed color coating and the others vary in a range of color coatings Λ. A color coating is a physical stimuli.

The second step is to specify the set of aspects $A(e)$ for every experience $e \in E$. Here, we take $A(e)$ to comprise: (a) the color qualities in e, that is, the experienced colors of

[9] *A variation $v : A(e) \to A(e')$ does not preserve a structure S with respect to relata $b_1, ..., b_m \in A(e)$* if and only if we have $R\big(b_1, ..., b_m\big) \neq R\big(v(b_1), ..., v(b_m)\big)$ if S is a relation R, or $v\big(f(b_1, ..., b_{m-1}\big) \neq f\big(v(b_1), ..., v(b_{m-1})\big)$ if S is a function f.
This negation agrees with the intuition because the definiendum already states part of the condition that follows, namely that $b_1, ..., b_m$ are relata of the structure S in $A(e)$, which implies that $(b_1, ..., b_m) \in R$ if S is a relation and that $f(b_1, ..., b_{m-1})$ exists in $A(e)$ if S is a function, meaning that the structure is satisfied before the variation.

the individual chips; (b) positional qualities of the color experiences, that is, which chip has which color; and (c) the experience of *relative similarity*. Relative similarity is an experience of one pair of aspects to be more, less, or equally similar to each other than another pair of aspects; here, the two pairs have to have one aspect—the reference—in common. In Figure 8.3.1A, for example, the color of the top left chip will, for many readers, be less similar to the reference chip than the color of the top right chip. An experience e in the set E may exhibit many other aspects as well. However, $A(e)$ only comprises those which are relevant for the construction at hand.

To pick out relative similarity more precisely, we let b_0, b_1 and b_2 denote the color aspects of the three chips in an experience e, where b_0 is the color aspect of the reference; see Figure 8.3.1B. For some experience e, it might be the case that the colors b_1 and b_0 are experienced as less similar to each other than the colors b_2 and b_0. In this case, the experience e has a relative similarity aspect in the above sense; we denote this "less-similar" relative similarity aspect by a. So, a is an aspect of e, and it is instantiated relative to b_1 and b_2. (To be precise, a is also relative to b_0. But since b_0 does not vary in E we can leave this implicit.)

Variations change one experience e into another experience e'. An example for a variation would be a swap of the coatings of the two non-reference chips, as in Figure 8.3.1C. Another example for a variation would be to change the coatings of both non-reference chips to some other coating in Λ, as in Figure 8.3.1D. Formally, variations are represented by mappings $v : A(e) \to A(e')$. In the first example, Figure 8.3.1C, the mapping is of the form $v(b_1) = b_2$ and $v(b_2) = b_1$, and $v(c) = c$ for all other aspects c, except for the relative similarity aspect a, which is discussed in detail below. In the second example, Figure 8.3.1D, the mapping is as in the first example but with $v(b_1) = b_3$ and $v(b_2) = b_4$.

The key question of this example is: Is there a mathematical structure of conscious experience which corresponds to relative similarity? To answer this question, we propose a mathematical structure and check whether this structure satisfies (MSC).

The words "less similar than" in the description of relative similarity already indicate that some order, in the mathematical sense of the word, might be involved. For reasons that will become clear below, we propose a strict partial order as mathematical structure. Our task in the remainder of this section is to show that this proposal indeed satisfies (MSC) with respect to experienced relative similarity. A strict partial order $(\mathbb{C}, <)$, consists of a set $\mathbb{C}$, which is the domain of the structure, and a binary relation '$<$' on $\mathbb{C}$. For all $x, y, z \in \mathbb{C}$, this binary relation has to satisfy the following axioms:

- *Irreflexivity*, meaning that there is no $x \in \mathbb{C}$ with $x < x$.
- *Asymmetry*, meaning that if $x < y$, then it is not the case that $y < x$.
- *Transitivity*, meaning that if $x < y$ and $y < z$, then also $x < z$.

In order to turn a strict partial order into a proposal for a mathematical structure of conscious experience, we need to specify how the set $\mathbb{C}$ and the relation $<$ relate to aspects of conscious experience. For the set $\mathbb{C}$ we choose the color qualities of the

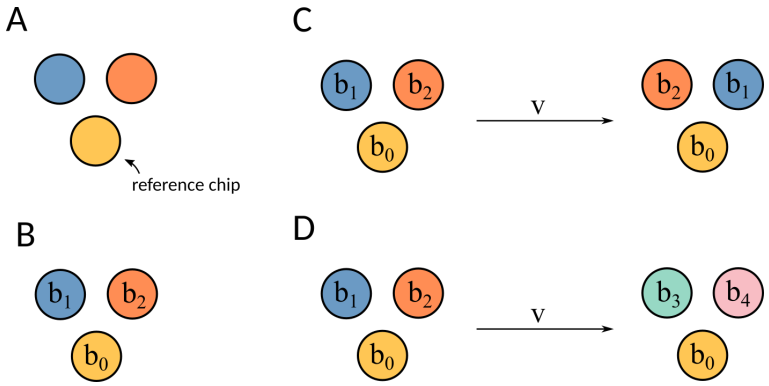


Figure 8.3.1.: To help explain the example of relative similarity, this figure illustrates experiences with color qualities and variations thereof. Subfigure A illustrates an experience of three color chips as well as the concept of *relative similarity*: many readers will experience the color of the top-left color chip to be *less similar* to the reference chip than the color of the top-right color chip. Subfigure B illustrates our notation for the color aspects corresponding to the color chips. Subfigures C and D illustrate variations v of experiences: a swap of two color aspects in C; and a replacement of two color aspects in D.

experiences in E, meaning that $\mathbb{C}$ now comprises the color qualities evoked by the coatings Λ of the chips we consider. For example, it contains what we have labelled b_0, b_1, b_2, b_3 and b_4 in Figure 8.3.1. For the relation, we define $b_i < b_j$ if and only if b_i is experienced as less similar to b_0 than b_j is to b_0. (Since relative similarity, as defined above, depends on the choice of reference b_0, it would be more precise to write $<_{b_0}$ instead of $<$. However, to simplify the notation, we keep the reference implicit.)

For this proposal to make sense, we first need to check whether the axioms are satisfied. If they were not satisfied, the proposal could still be a structure of conscious experience; but it wouldn't be a strict partial order. That's why the axioms are not explicitly mentioned in (MSC). Irreflexivity is satisfied because no color quality is less similar to the reference than itself. Asymmetry is satisfied because if a b_i is less similar to the reference than b_j, then b_j is not less similar to the reference than b_i.

The use of terms like 'less similar to' in natural language suggests that transitivity is also satisfied; it suggests that, if b_i is less similar to the reference than b_j and b_j is less similar to the reference than b_k, then b_i should be less similar to the reference than b_k. But it might very well be the case that natural language is not precise enough to describe its target domain. The use of natural language may be justified in simple cases, or even in a majority of cases, but whether or not transitivity holds for all $b_i, b_j, b_k \in \mathbb{C}$ is, ultimately, an empirical question. For the purpose of this example, we're going to assume that transitivity holds as well.

Having checked that the axioms hold—that is, that the proposal is indeed a strict partial order—we can proceed to check whether the structure is a mathematical structure

of conscious experience according to (MSC). Concerning Condition (S1), there is one domain $\mathbb{C}$ and it consists of color qualities, so this condition is satisfied. Therefore, only Condition (S2) remains to be checked.

We now show that the relative similarity aspect a, as defined above, is in fact an S-aspect, where S is the '$<$' relation on $\mathbb{C}$. That is, it is a $<$-aspect. To see that this is true, we have to show that a variation does not preserve $<$ with respect to relata b_1 and b_2 if and only if the variation changes a relative to b_1 and b_2.

Consider any variation $v : A(e) \to A(e')$ that does not preserve $<$ with respect to relata $b_1, b_2 \in A(e)$. Two aspects b_1 and b_2 are relata of $<$ if either $b_1 < b_2$ or $b_2 < b_1$. We focus on the first case as the other one follows from the first by renaming b_2 and b_1 in what follows. By definition of the $<$ relation, $b_1 < b_2$ means that b_1 is experienced as less similar to the reference than b_2. Therefore, there is also a relative similarity aspect $a \in A(e)$ as defined above. As explained in Section 8.2.3, there can be two ways in which the variation v might not preserve $<$. Either $v(b_1)$ or $v(b_2)$ are not defined, or, if they are defined, it is not the case that $v(b_1) < v(b_2)$. In the former case, there cannot be an a in $A(e')$ relative to $v(b_1)$ or $v(b_2)$, simply because the latter do not both exist. In the latter case, it follows from the definition of $<$ that $v(b_1)$ is not experienced as less similar to the reference than $v(b_2)$. So, there is no $a \in A(e')$ relative to $v(b_1)$ and $v(b_2)$. Hence, we may conclude that v changes a relative to b_1 and b_2.

For the opposite case, let $v : A(e) \to A(e')$ be a variation which preserves $<$ with respect to relata b_1 and b_2. As before, this implies that a is in $A(e)$ relative to b_1 and b_2. Because v preserves $<$, $v(b_1)$ and $v(b_2)$ both exist and we also have $v(b_1) < v(b_2)$. Applying the definition of $<$ then implies that a is also in $A(e')$ relative to $v(b_1)$ and $v(b_2)$. Hence v does not change a relative to b_1 and b_2.

Because in both of these cases, v was arbitrary, it follows that a is indeed a $<$-aspect. Therefore, Conditions (S1) and (S2) of (MSC) are both satisfied, and the strict partial order $(\mathbb{C}, <)$ is indeed a mathematical structure of conscious experience; it is the mathematical structure of relative similarity of color experiences with respect to b_0.

8.4. Phenomenal Unity and Topological Structure

Our final example concerns topological structure. Interestingly, this is intimately tied to phenomenal unity, the thesis that phenomenal states of a subject at a given time are unified (Bayne & Chalmers, 2003). Phenomenal unity gives rise to a mathematical structure of conscious experience.[10]

Recall that we have introduced the set $A(e)$ to denote aspects of the conscious experience e, where we have used the term 'aspect' as a placeholder for concepts like qualia, qualities, or (instantiated) phenomenal properties. Most examples of these concepts

[10] A connection between topology and phenomenal unity has already been conjectured in (Prentner, 2019), where an attempt was made to construct a topological space based on a binary relation that describes the "overlap" of mental objects. The construction only leads to the weaker notion of a pre-topology, but should be regarded as an important first step in this direction. For a summary of the formal construction, see (Kleiner, 2020b, Example 3.22).

are "independent" from the experience in which they occur; they could be experienced together with a largely different set of aspects in a different experience. Yet, experiences seem unified; their aspects are experienced as tied together in some essential way. This raises the question of what underlies this experience of the *unity of a conscious experience*? As we will see, somewhat surprisingly, the answer is: a topological structure of conscious experience.

Much has been written about the question of phenomenal unity in the literature, for example (Bayne, 2010; Bayne & Chalmers, 2003; Cleeremans & Frith, 2003; Mason, 2021; Prentner, 2019; Roelofs, 2016; Wiese, 2018), and in order to make use of some of the results, we assume that the term 'aspect' denotes an instantiated phenomenal property or quale. The set of aspects $A(e)$, then, comprises the phenomenal properties or qualia which are instantiated in the experience e, also called the *phenomenal states* of the experience e.[11] Our question, then, is what it means that "any set of phenomenal states of a subject at a time is phenomenally unified" (Bayne & Chalmers, 2003, p. 12).

There are various answers one might give to this question. A promising answer is the so-called *subsumptive unity thesis*, developed in (Bayne & Chalmers, 2003):

> "For any set of phenomenal states of a subject at a time, the subject has a phenomenal state that subsumes each of the states in that set." (Bayne & Chalmers, 2003, p. 20)

According to this thesis, what underlies the experience of the unity of a conscious experience is that for any set X of phenomenal states in the conscious experience, there is a further phenomenal state that subsumes each of the states in X. This phenomenal state characterizes what it is like to be in all of the states of X at once (Bayne & Chalmers, 2003, p. 20).

Put in terms of aspects, the subsumptive unity thesis says that for any set $X \subset A(e)$ of aspects of an experience, there is an additional aspect in $A(e)$ that subsumes the aspects in X. This aspect is the experience of what it is like to experience the aspects in X as part of one experience e together, the experience that they are *unified*, as we will say. Let us call this aspect the *phenomenal unity aspect* of X and denote it by a_X. It is instantiated relative to the elements of X.

Phenomenal unity gives rise to a mathematical structure of conscious experience. To see how, let us use the symbol $\mathcal{T}$ to denote a collection of subsets of $A(e)$, to be specified in more detail below. Every subset of $A(e)$ is a unary relation on $A(e)$,[12] and hence also on the set $\mathcal{A}$ that comprises all aspects of the experiences in E. Therefore, $(\mathcal{A}, \mathcal{T})$ is a mathematical structure; it has domain $\mathcal{A}$ and its structures are the unary relations in $\mathcal{T}$. As we show next, because of the subsumptive unity thesis, the mathematical structure $(\mathcal{A}, \mathcal{T})$ is a mathematical structure of conscious experience according to (MSC).

[11] A *phenomenal state* is an instantiation of a phenomenal property, or quale, by a subject at a given time. This instantiation constitutes part of the experience of the subject at the time. An experience e, in our terminology, is an experience of a subject at a given time. Hence, a phenomenal state is an instantiation of a phenomenal property, or quale, in an experience e.

[12] An m-ary relation on a set X is a subset R of X^m. Hence, a unary relation, where $m = 1$, is a subset of X.

Because $\mathcal{A}$ is the set of all aspects of E, Condition (S1) of (MSC) is satisfied. Therefore, only Condition (S2) remains to be checked. This condition is satisfied because for every set $X \in \mathcal{T}$, the phenomenal unity aspect a_X is an S-aspect for $S = X$; an X-aspect for short. To show that this is the case, we need to check that a variation does not preserve X with respect to relata $b_1, ..., b_m$ if and only if it changes a_X relative to $b_1, ..., b_m$. Let $v : A(e) \to A(e')$ be a variation that does not preserve X with respect to relata $b_1, ..., b_m$. The relata of the subset X are the elements of that subset. Therefore, we have $b_1, ..., b_m \in A(e)$, so that the subsumptive unity thesis implies that there is a phenomenal unity aspect a_X relative to the $b_1, ..., b_m$ in $A(e)$. The condition that v does not preserve X furthermore implies that either not all of the $v(b_i)$ exist or that at least one of them is not in the set X. Therefore, there is no phenomenal unity aspect a_X relative to $v(b_1), ..., v(b_m)$ in $A(e')$. Hence, the variation v changes a_X relative to $b_1, ..., b_m \in X$. Vice versa, let $v : A(e) \to A(e')$ be a variation which preserves X with respect to relata $b_1, ..., b_m$. This implies that a_X is instantiated relative to $b_1, ..., b_m$ in $A(e)$. The condition that v preserves X furthermore implies that $v(b_1), ..., v(b_m)$ exist, and that they are elements of X. Therefore, a_X is also instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$. This shows that the variation does not change a_X relative to $b_1, ..., b_m$. Thus, a_X is indeed an X-aspect. And because that is true for any $X \in \mathcal{T}$, $(\mathcal{A}, \mathcal{T})$ indeed satisfies Condition (S2) and hence (MSC).

The previous paragraph proves that, if the subsumptive unity thesis holds true for all sets X in $\mathcal{T}$, then $(\mathcal{A}, \mathcal{T})$ is indeed a mathematical structure of conscious experience. As we will explain next, this structure is intimately tied to a topological structure.

A topological structure $(M, \mathcal{T})$ consists of a set M and a collection $\mathcal{T}$ of subsets of M. The collection has to satisfy three axioms, and there are a few different ways of formulating these axioms. Here, we choose the formulation that corresponds to what is usually called 'closed sets'. The axioms are:

- The empty set $\emptyset$ and the whole set M are both in $\mathcal{T}$.
- The intersection of any collection of sets of $\mathcal{T}$ is also in $\mathcal{T}$.
- The union of any finite number of sets of $\mathcal{T}$ is also in $\mathcal{T}$.

Are these axioms satisfied by the structure $(\mathcal{A}, \mathcal{T})$ induced by phenomenal unity?

To answer this question, it is important to note that the subsumptive unity thesis does not provide a phenomenal unity aspect a_X for every subset of $\mathcal{A}$. It can only provide such an aspect for a set of aspects that are actually experienced together. That is, it can only provide such an aspect for a subset X of $A(e)$. Therefore, $\mathcal{T}$ is not the discrete topology introduced in Section 8.1. Second, it also cannot be the case that it provides a phenomenal unity aspect for every subset of $A(e)$. That's because then there would be an infinite regress: for every subset X of $A(e)$ there would be a new aspect a_X in $A(e)$, giving a new subset $X \cup \{a_X\}$ that would give a new phenomenal unity aspect $a_{X \cup \{a_X\}}$, and so forth. This problem is well-known in the literature (Bayne, 2005; Wiese, 2018). Rather, we take it, the quantifier 'any set' in the subsumptive unity thesis must be understood as 'any set of aspects that are experienced as being unified'. While it is arguably the case that the

whole set of aspects $A(e)$ of an experience is always experienced as unified–by which we mean: the whole set of aspects is experienced–, introspection suggests that we consciously experience only a select group of aspects as unified at a time.[13]

So, which sets of aspects do we experience as unified? While we cannot give a general answer to this question here, there is a special case where a sufficiently detailed specification can be given: the case of regions in visual experience. Here, 'regions' are sets of positions of the space that visually perceived objects occupy.[14] The positions in a region are experienced as unified. Therefore, the regions of visual experience are members of the collection $\mathcal{T}$ which is induced by phenomenal unity. Furthermore, they appear to satisfy the axioms of a topology as stated above: the whole set of positions in a visual experience is a region; it seems to be the case that intersections of regions in visual experience are also regions in visual experience; and it seems to be the case that the union of any two regions in visual experience is also a region in visual experience. For the empty set, no S-aspect of consciousness is required (there are no relata of the corresponding unary relation), so we may take the empty set to be a member of $\mathcal{T}$. Thus, all axioms of a topology are satisfied.

Therefore, if we take M to denote the position aspects of visual experiences, and choose $\mathcal{T}$ to comprise the regions of visual experience, then $(M, \mathcal{T})$ is indeed a topological structure. And, as shown above, it is a structure of conscious experience as defined in (MSC). We thus find that, because of the subsumptive unity thesis, this topological structure is indeed a mathematical structure of conscious experience; much like conjectured in (Tallon-Baudry, 2022), it is a topology of the visual content of subjective experience.

8.5. The Three Problems Revisited

In this section, we discuss how the new approach (MSC), which we have developed in Section 8.2.2, resolves the three problems discovered in Section 8.1.

Problem 1: Incompatible Structures

The first problem was that the condition (MDC), which has been applied in previous approaches, admits incompatible structures to conscious experience. Is this also true of (MSC)?

If two structures are incompatible, then there exists at least one automorphism of

[13] This solves the infinite regress problem because, arguably, we do not always experience the phenomenal unity aspects as unified with the sets they correspond to. So, there is not always a phenomenal unity aspect $a_{X \cup \{a_X\}}$ for the set that consists of a_X and X.

[14] It is also plausible to think that visual experiences do not contain positions as aspects, but only regions. However, assessing whether or not this is the case goes beyond the scope of this paper. Here, we assume that positions are aspects of visual experiences.

one structure that is not an automorphism of the other structure.[15] As we explain below, this condition implies that two incompatible structures cannot have an S-aspect in common. Therefore, it is not possible for two incompatible structures to pertain to conscious experience in the exact same way; so, (MSC) indeed resolves the problem of incompatible structures.

Let S and S' denote two incompatible structures with the same domains. Then, there is at least one automorphism of one structure that is not an automorphism of the other structure. Let us denote such an automorphism by v and assume that it is an automorphism of S but not of S'. Because v is not an automorphism of S', it follows that there is at least one set of relata $b_1, ..., b_m$ of S' in some $A(e)$, such that the variation $v : A(e) \rightarrow A(e)$ induced by the automorphism does not preserve S' with respect to these relata. On the other hand, because v is an automorphism of S, it follows that this variation preserves S with respect to $b_1, ..., b_m$. If an aspect a is an S'-aspect, then, applying the definition of S'-aspects, we find that the variation v needs to change it. In contrast, if an aspect a is an S-aspect, then, applying the definition of S-aspects, we find that the variation v must not change it; either because the $b_1, ..., b_m$ do not constitute relata of S, or because the variation v preserves S with respect to relata $b_1, ..., b_m$. Because an aspect cannot be both changed and not changed under a single variation, there cannot be an aspect a that is both an S-aspect and an S'-aspect.

Problem 2: Arbitrary Re-Definitions.

The definition (MSC) also resolves the problem of arbitrary re-definitions. That's because any re-definition changes the relations or functions of the respective structure, and therefore generates an own, independent condition for something to be an S-aspect of the redefined structure. Whether or not this new S-aspect is a part of conscious experience is a substantive question that depends on the actual experiences of the subject under consideration; it is not automatically the case.

Consider, as examples, the cases of rescaling a metric, which we have introduced in Section 8.1. If, per assumption, (M, d) were a structure of conscious experience, then for any relata $(b_1, b_2, d(b_1, b_2))$, the condition for d-aspects would have to be satisfied. Rescaling this to $(M, C \cdot d)$ generates a new condition because now, the relata to be considered are $(b_1, b_2, C \cdot d(b_1, b_2))$. These are different relata, and correspondingly, different experiences and different variations will enter the definition of a $C \cdot d$-aspect. The same is true for an $(f(a) + f(b)) \cdot d(a, b)$-aspect. Whether or not these structures satisfy (MSC) depends on the details of the conscious experiences under consideration; but they do not automatically satisfy (MSC) just because (M, d) does.

[15] Automorphisms are structure-preserving mappings from a structure to itself. Put in terms of the terminology we have introduced in Section 8.2.2, automorphisms are mappings v that map the domains of a structure to themselves. These mappings have to be bijective, and they have to preserve the structure, meaning that they have to satisfy (P1) for all elements of the domain in case of relations, and (P2) for elements of the domains in the case of functions.

Problem 3: Indifference to Consciousness.

The third problem is resolved, finally, because of the introduction of S-aspects in (MSC), which are a counterpart "in" conscious experience to the structure in the narrow sense of the term. S-aspects introduce a connection between functions or relations in a mathematical structure, on the one hand, and aspects (qualia, qualities, or phenomenal properties) of conscious experiences, on the other hand. Because S-aspects are part of the definition of (MSC), any application of (MSC) requires engaging with details of the conscious experiences of the subject under consideration; (MSC) is not indifferent to conscious experience in the sense of Problem 3 of Section 8.1.

Consider, for example, the two topological structures of Section 8.1. While (MDC) only required us to check whether the structures address aspects and satisfy the axioms, (MSC) also requires us to check whether there is an S-aspect in conscious experience that corresponds to the topological structures. As we have seen in Section 8.4, this involves a careful investigation of conscious experience and relies on intricate notions such as phenomenal unity.

8.6. Conclusion

In this article, we investigated mathematical structures and mathematical spaces of conscious experience. We were not concerned with questions of type or explicit form of these structures or spaces, but with the question of what it means to speak about mathematical structures or mathematical spaces of conscious experiences in the first place. We answer this question by providing a definition of what *mathematical structures of conscious experience* are. This definition provides a foundation for the construction, investigation and identification of concepts like phenomenal spaces, quality spaces, qualia spaces and Q-structures.

Our definition of mathematical structures of conscious experiences is grounded in a foundational understanding of mathematical structures and spaces as laid out by mathematical logic. And it is axiomatic in the sense that it can be applied to any conceptualization of conscious experiences, and any choice of aspects thereof (e.g. qualia, qualities, phenomenal properties, phenomenal distinctions), which satisfy the formal requirement that for every conscious experience there is a well-defined set of aspects.

Our definition rests on the notion of *variations*, which are changes of one conscious experience to another. Because variations can be induced introspectively (for example, as in Husserl's imaginary variations (Husserl, 1936)), stimulated in a laboratory by change of stimuli, or studied theoretically based on a proposed theory of consciousness, our definition constitutes a general method to identify and study structures of conscious experience.

The grounding of mathematical structures of conscious experiences proposed here is *methodologically neutral* in the sense that it can be combined with many methods, practices, and procedures that are used to investigate conscious experience, spanning empirical, analytical, and phenomenological research. Furthermore, it is *conceptually*

neutral in the sense that it can be applied to any conception of 'conscious experience' and 'aspects' thereof, as long as every conscious experience comes with a well-defined set of aspects. This includes common conceptions using qualities, qualia, or phenomenal properties, but also less common ideas based on atomistic conceptions of states of consciousness or phenomenal distinctions.

Our definition complements recent approaches that study quality spaces, qualia spaces, or phenomenal spaces, because it retains the abstract condition that these proposals apply—Condition (MDC) in Section 8.1—as a necessary part. This abstract condition is extended by our proposal, so as to avoid three problems that interfere with recent approaches, see Section 8.1.

In light of the increasing interest in using mathematical structures to model and represent conscious experiences in the scientific study of consciousness and philosophy of mind, the investigation of how to define and understand mathematical structures of conscious experience is important, in our view. This work contributes to this investigation. It highlights issues with previous ways of understanding structural claims and offers an improved conception that rests on meaningful desiderata. Hence, we hope, it contributes to building a foundation for structural research for both theory and experimental practice.

As a first application, and to illustrate our definition, we considered *relative similarity* and *topological spaces*. We found that relative similarity, which plays an important role in several constructions of quality spaces, is indeed a mathematical structure of conscious experience, see Section 8.3. Topological spaces also qualify as mathematical structures of conscious experience, but for a surprising reason: they are intimately related to phenomenal unity, see Section 8.4.

We view the results presented here as one further step in a long journey to investigate conscious experience mathematically. This step raises new questions and creates new opportunities, both of which can only be explored in an interdisciplinary manner. A new question, for example, is whether our result on mathematical structures might open new perspectives on measurements of consciousness (Irvine, 2013), as arguably promised by the Representational Theory of Measurement (Krantz, Luce, Suppes, & Tversky, 1971) whenever an axiomatic structure on a target domain is available. We hope that, ultimately, our result provides a basis for developing a common formal language to study consciousness across domains.

9. The Newman Problem of Consciousness Science

Johannes Kleiner

9.1. Introduction

The Newman problem is a fundamental problem for structural theories and structural assumptions throughout science. It was first raised by Newman (1928) in response to Russel's *The Analysis of Matter*, and concerns theories or assumptions which posit that:

> " 'There is a relation R such that the structure of the external world with reference to R is W.' " (Newman, 1928)[1]

Here, R denotes what would now be called the type of a structure. This could, to take a very simple example, be a partial order relation. W is a specification of such structure, meaning that it provides a set of mathematical objects—the elements that are to be related— and specifies which elements in that set are related by the binary relation.

The problem with such postulate is that "[a]ny collection of things can be organised so as to have the structure W, provided there are the right number of them. Hence the doctrine that only structure is known involves the doctrine that nothing can be known that is not logically deducible from the mere fact of existence, except ('theoretically') the number of constituting objects" (ibid.).

[1] "The world consists of objects, forming an aggregate whose structure with regard to a certain relation R is known, say W; but of the relation R nothing is known (or nothing need be assumed to be known) but its existence; that is, all we can say is, 'There is a [type of] relation R such that the structure of the external world with reference to R is W.' " (Newman, 1928).

It should not be immediately clear or self-evident why the antecedent of this statement is true—why any collection of things can be organised so as to have any structure W, provided there are the right number of them—, and neither why, if the antecedent is indeed true, it constitutes a problem. We explain why it is true, and why it does constitute a problem in Section 9.2.

Consciousness Science—the scientific investigation of conscious experience and its relation to the physical domain—is currently seeing early signs of a structural turn (Kleiner, 2024). As we will explain in Section 9.3, the Newman problem can undermine structural research and hence needs to be addressed for any theory-based structural research program to go ahead as intended.

The goal of this paper is to show that phenomenal spaces, and similar explications of the mathematical structure of conscious experience, do *not* suffer from the Newman problem, if the mathematical structure of conscious experience is understood in the right way. Nothing hinges on the particularities of consciousness here, other than that the methodology of structural claims that resolves the Newman problem was introduced in the context of consciousness. Therefore, we hope that this work might be of interest also to those who work on structural questions independently of consciousness.

9.1.1. Previous work

The Newman problem is almost 100 years old. Hence, it is no surprise that there is a large body of literature on the topic that discusses and clarifies the problem, as well as a host of different possible resolutions. We locate the work presented here in the landscape of existing resolutions in Section 9.6 and recommend (Frigg & Votsis, 2011) for an excellent review thereof.

In consciousness science, too, the problem has been discussed and resolutions have been proposed.

Lyre (2022) addresses the Newman problem in the context of a proposed relation between brain states and experiences called *Neurophenomenal Structuralism*. Here, the Newman problem threatens to undermine the claim that neural structures represent the structures of worldly states and processes. It constitutes a problem about what a subject can know about the world, so to speak. Lyre proposes a solution for the Newman problem that follows Russel's own answer to Newman (Russell, 2014), "that certain spatiotemporal [relations in the domain of worldly states and processes] do indeed carry over to [relations among neural states and processes]. We can indeed directly refer to certain spatiotemporal [relations in the domain of worldly states and processes]—or, in Russell's words, are 'directly acquainted' with them." (Lyre, 2022) That is the case, according to Lyre, because the sense organs encode the very spatiotemporal relations that govern external states, for example spatial changes or temporal differences.

Lyre's proposal targets the ramifications of the Newman problem for individual subjects and their epistemic or representational capacities. This paper, in contrast, is concerned with the abstract case of structural claims as part of scientific or philosophical theorizing.

Chalmers (2022) explains the Newman problem when applied to phenomenal consciousness in the context of his comparison of Carnap's logical construction of the world and Lewis's account of Humean supervenience. Chalmers endorses Carnap's resolution of the problem in terms of naturalness conditions (cf. Section 9.7.2), and concludes that "[b]ecause of Newman's problem, any construction system needs something extra-logical in the base" (Chalmers, 2022). Our result, albeit not spelled out in terms of the systems applied by Carnap or Lewis, challenges this claim.

The Newman problem also surfaces in the discussion of consciousness' potential intrinsic properties. Both (Seager, 2006) and (Brüntrup, 2011), for example, take the Newman problem to show that consciousness must be taken to exhibit intrinsic properties or intrinsic qualities. "It is very satisfying to see that the intrinsic nature argument is exactly what is required to avoid Newman's problem, and one would want it to be the case that both Russell and Eddington's deployment of consciousness as an intrinsic nature was explicitly directed at this issue" (Seager, 2006).

What our paper adds to this research is the proposal that if spaces and structure of conscious experience are understood in the right way from the start, no further resolution of Newman's problem is required.

9.1.2. Structure of this paper

After explaining Newman's problem in Section 9.2, we discuss it implications for consciousness science in Sections 9.3 and 9.4. Section 9.5 is devoted to explaining how a suitable definition of phenomenal spaces, and of mathematical structure of conscious experience more generally, avoids the Newman problem from the start. Section 9.6 explains how this generalizes to structural claims that do not target consciousness. Full mathematical details of the resulting general proposal are given in Appendix 9.A. Section 9.7 discusses the limits of the methodology we introduce, and concluding remarks are offered in Section 9.8.

9.2. The Newman Problem

Newman's problem arises because of what expressions like

$$\begin{gathered}\text{“the structure of the (...) world} \\ \text{with reference to } R \text{ is } W\text{”}\end{gathered} \tag{9.1}$$

(Newman, 1928) are traditionally taken to mean. For a mathematical structure W that comprises a domain C (the 'elements' of the structure) and a relation R, this traditional meaning consists of the following two conditions:

(D1) The elements of the domain C are properties of the world.

(D2) The relation R as specified by W exist.[2]

[2] This requirement is usually implicit in the requirement that the axioms of a structure (transitivity, reflexivity, anti-symmetry, for example, in case of a partial order), have to hold.

Because a mathematical relation R is a collection of tuples of elements of the underlying domain, Condition (D2) is actually stating that the tuples that constitute the relation R as specified by W exist.

Expressions like (9.1) are taken to be true if and only if (D1) and (D2) are true. Other than (D2), "of the relation R nothing is known (or nothing need be assumed to be known) but its existence" (ibid.). That is the content of (9.1) as traditionally conceived.

Let us consider, as an example, a partial order structure. Mathematically speaking, a partial order structure consists of a set of elements C, called the domain of the structure, on the one hand, and a binary relation R, on the other hand. The binary relation R is a subset of $C \times C$, meaning that it is a collection of tuples of the form (c_1, c_2), usually written as $c_1 \leq c_2$ in the case of partial orders. The fact that the partial order structure W consists of these two constituents is often expressed by writing $W = (C, R)$.

Condition (D1) then states that the elements of the domain C of the partial order (as specified by W) are properties of the world. Condition (D2) states that there exists a binary relation (viz. a collection of pairs of elements) that relates the elements as specified by W. The elements need to be arranged in tuples as specified by W for this condition to be true. This is what is means to say that a partial order structure W is a structure of the world, according to the traditional understanding of expressions like (9.1).

The problem with this understanding of (9.1) is that while elements in the domains of the structure are required to have referents in the world (a structure is a structure of the world only if there are properties as specified in the domain of W), this isn't true of the relation. The relation is not required to have a referent in the world. The condition on the relation is only exposed qua condition on the properties. In other words, the relation is not required to correspond to any concretum in the world, Condition (D2) only relates abstract formalism in W to abstract descriptions of the world. As a consequence, any abstract specification of structure in the world will satisfy (D2). This consequence is expressed by the following theorem, presented in (Frigg & Votsis, 2011; Ketland, 2004).

Theorem 9.2.1 (**Newman's Theorem**)**.** Let C be a collection of individuals and let W be a structure whose domain has the same cardinality as C. Then there exists a structure W_C whose domain is C and which is isomorphic to W.

That is to say, independently of whether the world actually comprises a relation R as specified by W, if the properties exist, one can simply define a suitable relation to render Condition (D2) true. "[G]iven any structure, if collection C has the same cardinality as that structure, then there is a system of relations definable over the members of C so that C has that structure. (...) [A]ll we have to do in order to define a relation is to put elements in ordered tuples and put these tuples together in sets, which we can always do as long as we have enough elements" (Frigg & Votsis, 2011). Condition (D2) is not itself depending on anything in the world over and above the dependence already established by (D1). This is the cause of the Newman problem.

There are a number of ways to resolve the Newman problem, cf. (Frigg & Votsis, 2011, Sec. 3.4) for an excellent discussion. When the problem is presented as above, the obvious route to a solution of the Newman problem is to ask whether one could not replace (D2) by a better condition, such that Theorem 9.2.1 ceases to apply. This route is

precisely the one we will take in Section 9.5, but first we will discuss why the Newman problem applies to consciousness science, and which ramifications it has for consciousness science.

9.3. The Newman Problem in Consciousness Science

Consciousness science is seeing early signs of what could be a structural turn. Virtually every field that is involved in consciousness science has started to employ mathematical spaces and mathematical structures as means to investigate, model, or measure the phenomenon.[3] In doing so, many different methodologies and ideas are applied, known under various different names, including quality spaces (A. Clark, 1993; Rosenthal, 2015; Lee, 2021), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021), qualia structure (Kawakita, Zeleznikow-Johnston, Takeda, et al., 2023; Tsuchiya et al., 2022), Q-spaces (Chalmers & McQueen, 2022; Lyre, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink et al., 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). All of these proposals attribute mathematical structure to conscious experiences, which is why we will use the term 'mathematical structure of conscious experience' as an umbrella term to refer to these and similar proposals.

In all of these proposals, there is a mathematical space or mathematical structure E that is claimed to describe, represent or model conscious experience. Modulo terminological choices, all of these proposals endorse some variant of the claim that

$$\text{“the structure of conscious experience is } E\text{”}. \tag{9.2}$$

In (Kleiner & Ludwig, 2024), we have analyzed those proposals that work with explicit conditions to assert such claims, cf (Kleiner & Ludwig, 2024, Sec. 1). Perhaps unsurprisingly, other than explicit statements of the axioms that a mathematical structure is required to satisfy, these are exactly Conditions (D1) and (D2), with 'properties of the world' replaced by 'properties of conscious experiences' or analogous constructs.

In (Kleiner & Ludwig, 2024), we use the term 'aspect' as a placeholder to denote concepts like qualia, qualities, instantiated phenomenal properties, phenomenal distinctions, or similar, that feature in claims of the form (9.2). For terminological simplicity, in this paper, we will work with the concept of *phenomenal properties*, which are properties of the phenomenal character of an experience, where 'phenomenal character' refers to what it is like for an organism to be that organism in a particular state (Nagel, 1974).[4]

[3] A list of references of current developments is given in Kleiner (2024).

[4] Philosophers often define an experience to be the instantiation of a phenomenal property by an experiencing subject, so that an experience is an event. The phenomenal character of an experience in this framework is what it is like for the subject to undergo such event. Cf. Nida-Rümelin (2018) for more details on and problems of this way of thinking. Those from a more formal context often tend to take

However, all we say below applies to other conceptual choices (qualia, qualities, phenomenal distinctions, etc.) as well.

In terms of phenomenal properties, we can formulate the conditions placed on structural claims in consciousness science as follows. Expressions like (9.2), where structure $E = (C, R)$ consists of domain C and relation R, are taken to be true if and only if

(C1) The elements of the domain C are phenomenal properties of conscious experiences.

(C2) The relation R as specified by E exists.[5]

To give an example of this in consciousness science, consider a metric space of color qualities. The requirement that a specification E of such space is a structure of conscious experience—a quality space, for short[6] —comprises, first, the condition that the points of the metric space are color qualities (being phenomenally presented with red, blue, etc.), and that the real numbers that describe the distances in a metric spaces are experienced degrees of similarity of such color qualities (being phenomenally presented with a degree of similarity). Color qualities are phenomenal properties, hence this requirement is Condition (C1). Second, for any two color qualities, there must be an experienced degree of similarity of those color qualities as described by the metric function.[7] This just means that any triple that consists of two color qualities and the corresponding degree of similarity as specified of the metric function must exist. This is Condition (C2). Only if both (C1) and (C2) are satisfied, is this an instance of a quality space.

Because conditions (C1) and (C2) are exactly analogous to conditions (D1) and (D2), the Newman problem applies to consciousness science in the exact same way as it applies in other domains. Explicitly, the Newman problem (Theorem 9.2.1) implies that any claim of the form (9.2) is empty, as far as the structural content is concerned. Nothing over and above the cardinality of the set of phenomenal properties is endorsed in a claim like (9.2). This has far-reaching consequences.

---

the term 'conscious experience' to refer directly to what it is like. In this paper, we will use the term 'phenomenal character' to denote what it is like, in the hope that this choice is the largest common denominator across fields and backgrounds.

[5] As in the case of general structural claims (Section 9.2), this requirement is usually implicit in the requirement that the axioms of a structure have to hold.

[6] For details on how quality spaces are constructed in consciousness science, cf. (Kleiner, 2024, Sec. 5).

[7] A metric space consists of two domains and one function. The domains are the set of points of the space and the real numbers. The metric function maps any two points to one real number. For simplicity, we have formulated (C2) in terms of relations only. Technically speaking, this condition can also be applied to functions because any function $d : C_1 \times C_2 \to \mathbb{R}$ is a unary relation on $C_1 \times C_2 \times \mathbb{R}$. We do think, however, that it is good to distinguish relations and functions in such contexts, and do so in (Kleiner & Ludwig, 2024).

9.4. Implications for Consciousness Science

The Newman problem has a number of ramifications in consciousness science. On the more obvious side of things are its ramifications for structuralist research programs. Less obvious, maybe, is that the Newman problem also undermines work on theories of consciousness.

9.4.1. Structuralist Research

Structuralist research programs in consciousness science come in one of two flavors. They either target the question of what can be known, scientifically or introspectively, about what it is like—what can be known about phenomenal character, in the terminology applied in this paper. Or they target the question of what phenomenal character actually is—in which sense it exists, so to speak. Following terminology of philosophy of science, we might designate the former as *epistemic phenomenal structural realism* (EPSR), and the latter as *ontic phenomenal structural realism* (OPSR). OPSR says that phenomenal structures are ontologically basic: non-structural features of phenomenal character, such as intrinsic qualities, do not in fact exist; only claims of the form of (9.2) can be true. EPSR is the view that all we can know about phenomenal character is its structure; only claims of the form (9.2) can be known. Cf. (Frigg & Votsis, 2011) for the corresponding distinction in philosophy of science. Therefore, if claims like (9.2) are in fact void over and above implications of cardinality, so are OPSR and EPSR. The Newman problem, if unresolved, undermines these research programs. This is well known, cf. e.g. (Lyre, 2022) or (Chalmers, 2023b).

9.4.2. Theories of Consciousness

What is less well known, maybe, is that the Newman problem also undermines theories of consciousness. Specifically, it undermines theories that address phenomenal structure, if those theories are intended to be applicable to non-human organisms or non-human systems more generally.

That is the case because for non-human systems, ostensive definitions of phenomenal structures fail. We cannot use language to pick out the referent of a structural claim like (9.2) in non-structural terms, either because non-human systems have no suitable language, or, in the case of LLMs, because they do not use language in the same way as we do. "The ostensive definition [only] explains the use—the meaning—of the word when the overall role of the word in a language is clear. Thus [only] if I know that someone means to explain a colour-word to me the ostensive definition 'That is called »sepia«' will help me to understand the word" (Wittgenstein, 1953). In human cases, we can get around purely structural claims like (9.2) by pointing out which phenomenal structure a phenomenal claim like (9.2) is intended to address. In non-human systems, because of the lack of shared meaning of language, this is not an option. The only thing we can

do is to specify the structure abstractly, as in (9.2), which is why the Newman problem applies in full force.

This is particularly evident in one of the mathematized theories of consciousness, Integrated Information Theory (IIT) (Albantakis et al., 2023; Oizumi et al., 2014). IIT comprises a carefully constructed algorithm that specifies, for any mathematical description of a system in a specific state, a complex mathematical structure called Φ-structure (cf. Kleiner and Tull (2021) for a structural exposition of IIT). The Φ-structure is the output of IIT's algorithm. In terms of the terminology applied here, it specifies the phenomenal character that a system is experiencing when it occupies the respective state. But IIT does not provide a phenomenal interpretation of this structure, it only provides the mathematics. This is a perfect example of (9.2).

To provide a Φ-structure is a substantial achievement of IIT. But if the Newman problem applies (which it must if (9.2) is understood as (C1) and (C2)), then IIT's structural claim is entirely void, over and above the cardinally of the elements in the structure.

This is related to what Chalmers (2023b) has called the Rosetta Stone problem of IIT: the problem of how to translate the mathematical structure that IIT proposes into phenomenological terms. If Newman applies, it follows that no such translation is possible, as the structural claim is void; IIT's structural claim can always be satisfied simply by defining the required structure over phenomenal properties.

The same applies to other theories of consciousness if they make structural phenomenal claims. Theories are prone to Newman's problem because they are supposed to stand on their own, they should be meaningful independently of ostensive human-language pointers that specify which structure is what in phenomenal character. It should suffice for a theory to specify the phenomenal structure of a system in terms of structural language; the relevant parts of phenomenal character should then be determined.

So how many theories address phenomenal structure and are intended to be applicable to non-human organisms or non-human systems? At present, only a small fraction of theories address phenomenal structure. Examples are IIT, mentioned above, as well as Expected Float Entropy Theory (Mason, 2021) and Rosenthal's quality-space version of higher order thought theory (Rosenthal, 2010). However, it can be argued that addressing phenomenal structure is inevitable once theories start addressing phenomenal character more faithfully than they presently do. Binary distinctions between whether a stimulus is being consciously perceived, or not, or whether a system is conscious at all, or not, might not suffice to explain phenomenal character faithfully (Kleiner, 2024). Furthermore, it can be argued that all theories of consciousness should be formulated in such a way that they can, in principle, be applied to non-human systems or organisms (Kanai & Fujisawa, 2024). This might be part of the desiderata for a theory to count as a meaningful theory of consciousness.

Therefore, the class of theories that will eventually come into the realm of the Newman problem is large. It looks like the Newman problem went by largely unnoticed, as far as scientific theories of consciousness are concerned, because most theories are not advanced enough at the present stage for them to introduce the tools that the Newman

problem vexes. But once they do, the Newman problem might well undermine much of the effort in constructing them, if unresolved.

9.5. Solving the Newman Problem of Consciousness Science

The obvious solution to Newman's problem, when presented as in Section 9.2, is to replace Condition (D2) resp. (C2) by another condition, so as to modify the meaning that expressions like (9.1) or (9.2) should have, in such a way that Newman's problem ceases to apply. This amounts to proposing alternative *definitions* of expressions like (9.1) or (9.2) that avoid Newman's problem; just like one proposes improved definitions of concepts like qualia or phenomenal consciousness in philosophy of mind to avoid problems the terms might otherwise face.

When improving (D2) resp. ((C2)), a new condition must remain compatible with the spirit of (9.1) resp. (9.2). The major constraints this raises is that the condition should also be formulated abstractly, and must only make use of exist quantifiers ('there exists ...'); no direct reference of properties of the world resp. phenomenal properties can be included. We now discuss three proposals of how this could be achieved. We employ the terminology of phenomenal properties, but the same points could be made with respect to properties of the world, as we explain in Section 9.6.

9.5.1. Higher-Order Phenomenal Properties

An immediate idea to improve Condition (C2) is to work with higher-order phenomenal properties. Phenomenal properties are properties of the phenomenal character of a conscious experience, and much like first-order phenomenal properties (presumably, for example, being phenomenally presented with red), there are higher-order properties (for example, being phenomenally presented with similarity of two shades of red).[8]

To improve (C2), one could simply add the condition that there exists a higher-order phenomenal property for every relation R in a structure E. That would amount to replaying (C2) by:

(C2′) The relation R as specified by E exists, and there is a higher-order phenomenal property.

The idea is that for every relation R, there is one higher order phenomenal property, and that no two relations can have the higher-order phenomenal property in common.

This condition would indeed resolve Newman's problem because the simple existence of a structure with phenomenal properties as its domain is not sufficient any more to satisfy (C2′). Rather, there must be a phenomenal property (or something in the world, in

[8] In (Kleiner & Ludwig, 2024), we have called these 'structural properties', but this choice of terminology might not be ideal as it suggests that these properties already have some structure in the mathematical sense of the term. This is not the case. Rather, they only have *arity* (the number of lower-order properties they are instantiated relative to, cf. Section 9.5.2 below).

Newman's terms). This is an additional requirement whose satisfaction does not follow from Theorem 9.2.1.

However, (C2′) is not a suitable proposal because the structural phenomenal property that needs to exist has nothing to do with the relation R as specified by the mathematical structure E. The condition does not pin down the relation in any significant sense, over and above the requirement that the number of relations that exist is smaller than the number of higher-order properties. It's not enough to just require *some* phenomenal property to exist.

9.5.2. Arity

In order to remedy the problem of (C2′) that the mathematical structure of E has nothing to do with the higher-order phenomenal property that is required to exist, we have to expand the requirements placed on the higher-order phenomenal property.

While higher-order phenomenal properties do not have, or cannot be taken to have in this context, a mathematical structure that one can simply reference, they do exhibit a feature that in mathematics is called *arity*, and in philosophy may also be called *adicity*. It is the number of lower-order properties the higher-order property is instantiated relative to. For example, if the higher-order property is being phenomenally presented with similarity of two different shades of red, it has an arity of 2.

Relations in the mathematical sense of the term also have arity. It is the number of "slots" in the relation, or in other words, the number of elements that every tuple in the relation comprises. A binary relation, for example, has arity 2 because its tuples are pairs of elements. A relation of arity n comprises n-tuples, each of which consists of a list of n elements of the domain. Making use of this fact, we could modify (C2′) to read:

(C2″) The relation R as specified by E exists, and there is a higher-order phenomenal property that has the same arity as R.

This is an improvement over (C2′) because now the phenomenal property cannot be arbitrary any more.

However, (C2″) still fails because there are vastly different relations of the same arity. Arity characterizes a relation to some extent, but it still leaves most details of a relation unspecified.

What is needed to arrive at a satisfying condition is some way of characterizing a relation's mathematical form, that can also be interpreted in terms of phenomenal character.

9.5.3. Automorphisms

One way to characterize a mathematical structure in full (up to a certain point, cf. Section 9.7) is given by its *automorphism group*. Automorphisms are functions that map every element from a domain of a structure to another element of the domain. The mapping has to be one-to-one (implying that it has to be invertible), and has to preserve the

relations (and functions) defined over a structure. If a structure consists of one domain C and one binary relation R, for example, this mapping takes the form

$$f : C \to C \,,$$

and the requirement that it preserves the relation is formaly stated as

$$R(c_1, c_2) = R\left(f(c_1), f(c_2)\right) \tag{9.3}$$

for all $c_1, c_2 \in C$.[9] Automorphisms form a group because they are invertible, and because any two automorphisms can be concatenated to give a new automorphism.

Automorphisms are intriguing objects in the current context because, once a domain is specified (qua Condition (C1)), a set of automorphism can be specified as a set of functions $\{f_1 : C \to C,\ f_2 : C \to C,\ ....\}$. Neither the relation R, nor the tuples that constitute the relations, have to be specified when specifying the functions in the set.

Of course, if one would only specify a set of automorphisms, the Newman problem would apply just as well. They are formal objects and hence always exist, if the domain contains enough elements. What is needed, in addition, is a link between automorphisms and phenomenal properties. Such a link can be provided, as we now explain.

Let us consider an arbitrary function (also called 'mapping') $f : C \to C$, where C is a domain of a structure that satisfies (C1). An arbitrary mapping can or cannot be an automorphism of a structure E, depending on whether it satisfies the definition of an automorphism, or not—that is to say, depending on what the structure E is, and depending on how elements are mapped by the function. If a function is an automorphism, one often says that it "preserves" the structure. If it is not an automorphism, one says that it does "not preserve" the structure. Those are abstract statements in the domain of mathematics. (Cf. Definition 9.A.2 in Appendix 9.A for formal details.)

But in cases where a domain C satisfies (C1), functions $f : C \to C$ can also be understood as something concrete: they describe how phenomenal properties change. To give a very simple example: if a subject has an experience of seeing red, and that changes to an experience of seeing blue, this can be described as a (partial) function that maps from phenomenal properties to phenomenal properties; it maps being phenomenally presented with red to being phenomenally presented with blue. Such a variation of phenomenal properties must, in turn, be understood as a variation of the underlying experience whose phenomenal properties are at issue. Variations of experiences are changes from one experience to another, and for every such change, there is a corresponding variation of (instantiated) phenomenal properties.

Because functions can be interpreted in both abstract and concrete terms, they provide the link between the abstract domain of mathematics and the concrete domain of conscious experiences that is needed to amend Condition (C2″). They allow us to express the requirement that the higher-order phenomenal property in (C2′) mirror the structure E in terms of behavior of variations as follows: a higher-order phenomenal property

[9] We write an equal sign here for notational simplicity. The formally correct statement would be $R(c_1, c_2) \Leftrightarrow R\left(f(c_1), f(c_2)\right)$ for all $c_1, c_2 \in C$.

must behave as the structure does under variations. This means that the higher-order phenomenal property must prevail in phenomenal character if the mapping between first-order properties induced by a variation preserves the structure; and it must disappear if the mapping between the first-order properties induced by a variation does not preserve the structure. We will say that in former case, the variation "preserves" the higher-order phenomenal property, whereas in the latter case it "does not preserve" the higher-order phenomenal property (cf. Definition 9.A.1 in Appendix 9.A for formal details).

We call a higher-order property that satisfies this requirement for a relation R a phenomenal R-property. In concise terms:

(SP) A phenomenal property p is an R-property iff any variation that preserves the relation R preserves p.

We note that this definition is only meaningful if (C1) holds; (C1) provides the "baseline correspondence" between mathematical structure and phenomenal properties that allows to make sense of variations both in terms of changes of phenomenal properties and automorphisms.

We can thus present a suitable extension of (C2) as:

(C2‴) The relation R as specified by E exists, and there is a corresponding higher-order phenomenal R-property.

This is the meaning/definition of structural claims like (9.2) we have arrived at, for independent reasons, in (Kleiner & Ludwig, 2024) as well. It makes use of variations which form an important part of earlier proposals of how to define spaces of conscious experiences, for example Rosenthal (2015), and it retains the original Condition (C2): since (C2‴) implies (C2), Condition (C2) is a necessary part of Condition (C2‴).

Condition (C2‴) resolves the Newman problem because the mere existence of some structure is not sufficient to satisfy the condition. The condition requires that there is a phenomenal property of the right sort. This is a requirement whose satisfaction does not follow from Theorem 9.2.1. The condition furthermore leaves no freedom for the relation to vary while the property is fixed, as (C2′) and (C2″) did. Hence it is, as far as we can see now, a viable solution of the Newman problem of consciousness science.

9.6. A general solution?

In the previous section, we have shown how the Newman problem of consciousness science can be resolved by providing a more careful definition of what structural claims are taken to be. Here, we discuss whether this affords a solution of the Newman problem independently of consciousness.

Before we embark on this discussion, we would like to mention that there are several viable solutions of the Newman problem already, discussed in detail in (Frigg & Votsis, 2011). A review of these solutions would go beyond the scope of this paper, but suffice

it to say that the solution presented here might be a case of the 'Real vs. Fictional Relations' class of solutions that attempt to distinguish real relations in the world from those that are merely defined (called 'fictional' by (Newman, 1928)).

The viability of our solution of the Newman problem in consciousness science as a solution of the general Newman problem depends on whether the ingredients we have made use of also exist in the general setting of the Newman problem. These are higher-order phenomenal properties, and the concept of variations.

The notion of higher-order phenomenal property easily carries over to any context in which a structural claim like (9.1) is made. There are higher-order properties of the world in a similar sense as there are higher-order phenomenal properties.[10]

The case of variations is more difficult. The crucial property of variations that enables definition (C2‴) is that the variation changes both the first-order property, and the higher-order property. Can we make sense of such variations? And if so, what defines the variations that exist as compared to those that do not.

In a context like (9.1), where the reference of a structural claim is "the world", one could make sense of variations in terms of possible world semantics as used in modal logic, cf. e.g. (D. Lewis, 1986). Specifically, one could consider the set of all nomologically possible worlds—the set of worlds that are compatible with the laws of nature, that is—and then define a variation simply to be a map from one nomological possible world w_1 (e.g., the actual world) and all its properties to another nomological world w_2 and all its properties. Such a variation preserves a (possibly higher-order) property if and only if it is present before and after the variation, meaning if it is a property of both w_1 and w_2, and it preserves a structure S if and only if it is an automorphism of S, where the latter definition makes use of (D1).[11]

This gives rise to the following exposition of structural claims like (9.1). The notion of R-property is defined as:

(SP) A property p is a R-property iff any variation that preserves the structure R preserves p.

Claim (9.1) is true if an only if the following two conditions are true:

(D1′) The elements of the domain C of are properties of the world.

(D2′) The relation R as specified by W exists, and there is a higher-order R-property.

[10] If properties are conceived of as properties of things, one might want to distinguish the concept of higher-order properties from the concept of relational properties. Relational properties are properties between things. There can be both first-order relational properties, and higher-order relational properties. Properties which are properties only of one thing are called monadic properties, and there are both first-order and higher-order monadic properties, the latter of which are properties of properties of one thing. According to this conception of properties, our proposal below could be defined in terms of either relational or higher-order properties, or both; what matters is that the properties in question have arity, also called adicity. I would like to thank Andrew Lee for pointing this out.

[11] Because properties can disappear from w_1, as in the case of consciousness, mappings must be understood as partial functions. Because they need not be surjective, properties can appear in moving to w_2. More details on such mathematical subtleties are given in the appendix, and in (Kleiner & Ludwig, 2024).

While abstract at first, this condition is highly compatible with physical sciences, because nomologically possible worlds should not be understood as atomistic entities. Rather, the set of nomologically possible worlds is intimately connected with initial conditions of natural laws, and a fortiori with repeated experiments. Structural claims so defined can be assessed empirically by considering "chunks" of the actual world in scientific experiments, and by studying how these chunks behave as time or other parameters vary.

An alternative to this approach would be to take properties in the world to be attached to objects, or groups of objects, in the world, and to consider variation of such (groups of) objects. This would also provide a suitable concept because a variation of a (group of) objects would vary both first-order and higher-order properties of the object or group. A definition of this kind would be of advantage because it would be more intuitive as the above. However, it would not naturally align with the foundations of physics, where existence of individual objects (rather than just one global field with particles as modes or excitations thereof) in an intuitive sense is somewhat contested. Still, it might be a viable option, and might actually correspond to the above definition if possible worlds are conceptualized in the appropriate way.

We provide a full formal exposition of our proposal in Appendix 9.A, and discuss a limitation of our approach in Section 9.7. The consequences of this limitation are, on our view, what ultimately determines the viability of our proposal for purposes of solving the general Newman problem. In the next section, we explain how our proposal relates to the Newman problem when expressed in terms of Ramsey sentences.

9.6.1. Ramsey sentence formulation of the Newman problem

The Newman problem is often stated in terms of Ramsey sentences, introduced by Carnap (D. Lewis, 1970). In a nutshell, for any theory T that contains observational predicates Q_i and non-observational predicates P_i, one can first form a logical conjunction of all of a theory's postulates/axioms/rules to write the theory as a single formal sentence that is usually denoted as (Frigg & Votsis, 2011)

$$T\left(P_1, ..., P_m, Q_1, ..., Q_n\right) . \tag{9.4}$$

The Ramsey sentence of such theory is the result of replacing all non-observational predicates P_i by variables, which we denote as X_i, and adding an existential quantifier over these variables, denoted by '$\exists$' to the sentence.[12] This gives the theory's Ramsey sentence T_R,

$$\exists X_1 ... \exists X_m \, T\left(X_1, ..., Q_n\right) . \tag{9.5}$$

A Ramsey sentence encodes a theory's full empirical content. Because the predicates P_i are non-observational predicates, they do not have observational consequences over and above their mere existence and role in the theory T. Therefore, a theory and its

[12] This is an instance of quantification over predicates, which presumes second-order logic.

Ramsey sentence have the same observational consequences. Furthermore, the Ramsey sentence (9.5) follows logically from (9.4). Cf. (Frigg & Votsis, 2011, Sec. 3.3) for more details.

Making use of Ramsey sentences, the Newman problem can be stated as the following theorem. Here, a model of a theory is u-*cardinality correct* if it has the same cardinality as the unobservable predicates of a theory, and *empirically correct* if its empirical substructure is isomorphic to the empirical substructure of the target domain (Frigg & Votsis, 2011).

Theorem 9.6.1 (**Cardinality Theorem**)**.** The Ramsey Sentence of theory T is true if, and only if, T has a model S (i.e. $S \models T$) which is u-cardinality correct and empirically correct.

This theorem establishes that "all we can infer from the truth of [a theory's Ramsey sentence] T_R about the unobservable world is a claim about its cardinality" (ibid.), and that "any claim the [Ramsey sentence] may make about the existence of unobservable relations or their formal properties is automatically true (or 'trivially' true, as the point is often put)" (ibid.).

How does our proposal deal with the Ramsey sentence formulation of the Newman problem?

Our proposal amounts to a redefinition of the truth-condition of structural claims. According to the received view of such truth-conditions, a second-order predicate that expresses a structural claim is true iff Conditions (D1) and (D2) are true. According to our proposal, a second-order predicate that a expresses a structural claim is true iff Conditions (D1′) and (D2′) are true.

This changes the implications of the existential quantifiers in (9.5). They do not assert that there exists structure in the world that satisfies (D1) and (D2), but rather that there exists structure in the world that satisfies (D1′) and (D2′).

As a consequences, the right-to-left direction of Theorem 9.6.1 breaks down. While it is still true that the truth of a Ramsey sentence of a theory T implies that there is a model which is u-cardinality correct and empirically correct (Condition (D2) is still a necessary part of Condition (D2′); this is the left-to-right direction of the theorem), the opposite direction fails to hold: it is not the case that any model which is u-cardinality correct and empirically correct implies the truth of the Ramsey sentence, because it also needs to satisfy the R-property condition in (D2′).

As a consequence, with the improved understanding of structural claims that we have proposed above, it ceases to be true that "any claim the RS may make about the existence of unobservable relations or their formal properties is automatically true (or 'trivially' true, as the point is often put)" (ibid.).

9.7. Objections

In this section, we would like to address one objection to, and one fundamental worry of, our proposal.

9.7.1. Reconstructing structure

The fundamental worry concerns the question of just how much of a mathematical structure can be identified (or "reconstructed") from its automorphism group.

Consider again the three proposals we have made in Section 9.5. Starting from the fundamental idea to add existential quantifiers of higher-order properties in (C2′), we have subsequently expanded the condition so as to limit the number of mathematical structures that can be associated with a given higher-order property. While Condition (C2′) did not put any constraint on how the structure relates to a higher-order property, Condition (C2″) required arity to match up, and Condition (C2‴) required the variations that constitute a structure's automorphism group to match up with the variations that preserve the higher-order property.

The problem we discuss here is that while it is true that an automorphism in general characterizes a mathematical structure in full, it does not do so in extreme situations. Automorphisms "min out" at some point. Once the automorphism group is trivial, it remains trivial even if more structure is added, as we now explain.

Consider a mathematical structure W that consists of a domain C and relation R, where the relation R allows to individuate every element of C uniquely based on relational information alone. This is the case for graphs, for example, if every node of a graph has a unique number of edges that connect to that node, called the degree of that node. The automorphism group of such structure contains only the identity mapping, for every other mapping would not be able to preserve the edge relation (cf. (9.3)). In this case, the automorphism group of the structure is called 'trivial'.

Trivial automorphism groups constitute a problem for our proposal because once the automorphism group is trivial, automorphisms fail to track any further changes to structure that preserve triviality. If, for example, a further edge is added to a graph, while preserving the condition that every node has a unique degree, then the automorphism is trivial before and after the change in structure. It can neither track, nor be used to reconstruct, the difference in structure. Put more abstractly, different relations that are defined over a given (fixed) set of elements can all have the same trivial automorphism group. The condition for there to be an R-property is the same for any relation R that satisfies (D1′) for a given set C and whose automorphism group is trivial. This problem of automorphism-based criteria to distinguish structure is well-known in the structural parsimony debate in philosophy of physics (T. W. Barrett, Manchak, & Weatherall, 2023).

There are two different responses one can give to this problem, and both apply.

First, one could argue that this problem indicates that Condition (D2′) can still be improved. Maybe some more advanced math could be used to resolve structure via automorphisms even if the automorphism group of a structure is trivial. Local automorphisms and sheaves come to mind. Or maybe there is an entirely different way of formulating a condition that replaces (D2). Both are viable options to explore in further research.

Second, one could argue that the problem is not actually detrimental to the proposal, because such relations cannot satisfy both Conditions (D1′) and (D2′).

To see why this is the case, we first emphasize that the problem we describe here only

applies if individual relations already imply that the automorphism group of a structure is trivial. That is because every relation R is required to have a corresponding R-property. If a structure contains more than one relation, and the entirety of them render the automorphism group trivial, all is well.

Consider, therefore, a single relation R that precludes non-trivial automorphisms. For this relation to satisfy (D2′), there needs to be an R-property p as described by (SP). This yields two conditions: First, any variation that preserves R needs to preserve p. And second, any variation that does not preserve R must not preserve p. The automorphism group being trivial implies that only the second case applies, so that any variation whatsoever must not preserve p. The mathematical formalism of our proposal implies that if there is to be an R-property p, any world which instantiates the elements in one tuple of R must instantiate p. Because the relation has trivial automorphism, there must at least be two tuples in the relation. Therefore, we must at least have two worlds that instantiate p. But any variation from one of these worlds to the other of these worlds preserves p, as it is instantiated both in the source and target world of the variation. This violates the condition that there is no variation that preserves p, and hence there cannot be an R-property for a relation that has trivial automorphism.[13]

The formal arguments behind this reasoning are provided rigorously in Appendix 9.B (cf. Lemma 9.B.2). In summary, the mathematics of our proposal simply deny relations that induce trivial automorphism groups the status of viable objects of a structural claim. This is aligned with the idea that a mathematical structure is only meaningful to the extent that it can be probed by variations.

9.7.2. Does naturalness suffice?

One way to resolve the Newman problem is to assume that "only natural relations should be taken into account when pondering the structure of the world; we need not, strictly speaking, deny that the world instantiates (...) any relation compatible with its cardinality, but we submit that only natural relations are taken into account when it comes to assessing the claims of a theory" (Frigg & Votsis, 2011). This idea was introduced by D. Lewis (1983), and is the solution endorsed by Chalmers (2022) in the context of consciousness science, cf. Section 9.1.1.

Given that our proposal in (D2′) introduces a technical term, one could object that the solution in terms of naturalness is preferable, simply because it is a simpler solution. Is this so?

The solution terms of naturalness amounts to reinterpreting the 'exists' quantifier in (D2). Instead of an abstract existential claim, it would have to be interpreted as quantifying over natural *relations*. If there is a natural relation R as specified by W, then (D2) is true. If not, (D2) is false.

This resolution of the Newman problem is problematic, cf. (Frigg & Votsis, 2011, Sec. 3.4.1(b)). One problem is that what counts as a natural kind might change as science progresses, cf. (Melia & Saatsi, 2006). Another problem is that distinguishing nat-

[13] For a formal proof of this claim, cf. Lemma 9.B.2 of Appendix 9.B.

ural kinds from non-natural kinds might require non-structural language of the world to begin with, cf. (Psillos, 2005).

But more fundamentally, even, for this solution to work and be applicable one has to presume that the world is mathematical, and that mathematical terms refer "just like that". One has to presume that it is meaningful to say there is a natural *relation*, where 'relation' is used in the mathematical sense of the term.

This is a substantive claim, and at least when it comes to phenomenal character, there are good reasons to think it is wrong. Conscious experiences do not come with mathematical structure in any meaningful way. Phenomenal character isn't experienced as a metric space, for example. There are experiences, and mathematical formalism is useful to describe or represent experiences. To say that conscious experiences have a mathematical structure is a way of describing them, not part of what they are naturally given as. Similarly, natural kinds (or related concepts) might not constitute mathematical structure "just like this".[14]

If this is true, then the naturalness solution is in fact solution (C2′), where 'phenomenal properties' are replaced by 'natural properties'. The reasons for rejecting this proposal in favor of (C2‴) apply mutatis mutandis to natural properties. As a result, the naturalness solution might not get around the introduction of the technical terms in (D2′). It might simply amount to (D2′) formulated with (higher-order) natural properties. An important change in cases where inflationary conceptions of properties are involved, but otherwise not substantially different.

9.8. Conclusion

We have considered how the Newman problem applies to consciousness science, and shown that it threatens to undermine structural research and structural theories that target conscious experience.

The problem resides in the particular understanding of structural claims that is presumed when discussing phenomenal spaces, quality spaces, qualia spaces, experience spaces and the like. If unresolved, research that subsumes this understanding is inherently limited and prone to errors. As far as theoretical work is concerned, use of such spaces simply doesn't make sense with the usual subsumption of structural claims.

However, when one adopts are more careful definition of structural claims, the Newman problem ceases to apply. The upshot of our discussion, framed in terms of phenomenal properties for simplicity, is that if structural claims like "the structure of conscious experience is E" are taken to be true if and only if the following two conditions hold, the Newman problem ceases to apply.

(C1) The elements of the domain C of are phenomenal properties of conscious experiences,

[14] There is also a worry of circularity here, if in order to be able say that the world has some mathematical structure, one needs to be able to say that natural kinds have such structure.

(C2) The relation R as specified by E exists, and there is a corresponding higher-order phenomenal R-property,

Here, a phenomenal property p is an R-property iff any variation that preserves the relation R preserves the phenomenal property p.

What distinguished our proposal from previous approaches is only the inclusion of R-properties in (C2). This suffices to resolve the Newman problem and the negative consequences that otherwise apply. While abstract at first, this proposal is straightforwardly applied to existing cases, and in fact builds on previous definitions of quality spaces, as explained in (Kleiner & Ludwig, 2024).

For readers with a broader background in philosophy, we have presented our proposal in general, consciousness-independent terms in Section 9.6 and Appendix 9.A. Whether or not this proposal is helpful in the general discussion of the Newman problem, and whether it can be applied to domains other than consciousness, is an open question.

Acknowledgments

I would like to thank Tim Ludwig and Andrew Lee for discussions on the topic and feedback on earlier versions of the manuscript, as well as the NYU Center for Mind, Brain, and Consciousness for hosting me while working on this manuscript.

Appendix

9.A. Full Definitions

Here, we provide formal details of our proposal of the general Newman problem as presented in Section 9.6. We denote a mathematical structure by $\mathbb{S}$. It is a tuple

$$\mathbb{S} = \big((\mathcal{A}_i)_{i\in I}, (S_j)_{j\in J}\big)$$

of domains $\mathcal{A}_i$ and functions or relations S_j. We denote the class of nomologically possible worlds by W and the properties of a world $w \in W$ by $\mathcal{A}(w)$. For expository reasons, we define the class of all properties of all worlds by

$$\mathcal{A} = \bigcup_{w\in W} A(w)\,. \tag{9.6}$$

A *variation* of a world w changes w into another world w'. Because worlds have structure, there may be various different ways to go from w to w'.[15] Therefore, in addition to specifying w and w', a variation is a partial mapping

$$v : A(w) \to A(w')\,.$$

This mapping describes how properties of the world w are replaced or reshuffled by the variation. A mapping which is not surjective, meaning that it does not map to all properties in $A(w')$, makes room for appearance of new properties of w'. A mapping which is partial, meaning that it does not specify a target for every property in $A(w)$, makes room for properties to disappear.

Higher-order properties are properties that are instantiated relative to other properties. If a property a requires other properties for its instantiation, we will say that the aspect a *is instantiated relative to* properties $b_1, ..., b_m$, or simply that a *is relative to* $b_1, ..., b_m$. Higher-order properties are the building blocks for our proposal to define structual claims like (9.1).

[15] To illustrate this point, consider the following example, provided in (Kleiner & Ludwig, 2024). Let v and v' be mappings that map the numbers 1, 2, and 3 to the numbers 2, 4, and 6. The mapping v is the multiplication of every number by 2, meaning that we have $v(1) = 2$, $v(2) = 4$, $v(3) = 6$. The mapping v', on the other hand, is defined by $v(1) = 6$, $v(2) = 2$, $v(3) = 4$. If we only cared about the *sets* of elements that these mappings connect, the mappings would be equivalent: there is no difference between the set $\{2, 4, 6\}$, which is the image of v, and $\{6, 2, 4\}$, which is the image of v'. If, however, we care about the *structure* of the elements of the sets–in this case, the *ordering* of numbers–, then there is a difference. While $2 \leq 4 \leq 6$, it is not the case that $6 \leq 2 \leq 4$. Because we care about the order of the elements, we need to say which element goes where.

Definition 9.A.1. A variation $v : A(w) \to A(w')$ **does not preserve a property** $a \in A(w)$ *relative to* $b_1, ..., b_m \in A(w)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(w)$, but a is not instantiated relative to $v(b_1), ..., v(b_m)$ in $A(w')$.[16]

In the case where $a \in A(w)$ is not a higher-order property, this definition reduces to the simple condition that $a \in A(w)$ but $a \notin A(w')$. The negation of the definition is also as intuitively expected: the property is present both in the source and in the target.[17]

For applications it is important to understand that this definition can fail to apply in two ways. First, it can fail because there is no a in $A(w')$ which is instantiated relative to $v(b_1), ..., v(b_m)$. This, in turn, can be the case either because there is no a in $A(w')$ at all, or because there is an a in $A(w')$ but it is instantiated relative to other aspects. Second, it can fail because one or more of the $v(b_1), ..., v(b_m)$ do not exist. The second case is possible because v is a *partial* mapping, which means aspects can disappear.

We use the term *relata* to designate those elements of a domain that are related by a structure. In the case where S is a relation R on a domain $\mathcal{A}$ and has arity m, these are the elements of the m-tuples $(b_1, ..., b_m) \in R$. In the case where S is a function $f : \mathcal{A}_1 \times ... \times \mathcal{A}_{m-1} \to \mathcal{A}_m$, the relata are the elements of the m-tuples $(b_1, ..., b_{m-1}, b_m)$ where $b_m = f(b_1, ..., b_{m-1})$, and where the other b_i range over their whole domains. For notational simplicity, we write $b_1, ..., b_m$ instead of $(b_1, ..., b_m)$ when designating relata in what follows.

Definition 9.A.2. A variation $v : A(w) \to A(w')$ **preserves a structure** S *with respect to relata* $b_1, ..., b_m \in A(w)$ if and only if we have

(P1) $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ if S is a relation R, or[18]

(P2) $v\big(f(b_1, ..., b_{m-1})\big) = f\big(v(b_1), ..., v(b_{m-1})\big)$ if S is a function f.

As in the previous case, the negation of this definition is exactly what is intuitively expected: a variation does not preserve the structure if and only if the structure is satisfied before the variation, but not satisfied after the variation.[19]

---

[16] In (Kleiner & Ludwig, 2024), we use the term 'changes' rather than 'does not preserve'. In hindsight, we think it is easier to speak of preservation too in this case.

[17] Because the definiendum already includes the first part of the condition, the negation is as follows:
A variation $v : A(w) \to A(w')$ *preserves a property* $a \in A(w)$ *relative to* $b_1, ..., b_m \in A(w)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(w)$ and a is also instantiated relative to $v(b_1), ..., v(b_m)$ in $A(w')$.

[18] For notational simplicity, we write $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ instead of $R\big(b_1, ..., b_m\big) \Leftrightarrow R\big(v(b_1), ..., v(b_m)\big)$.

[19] *A variation* $v : A(w) \to A(w')$ *does not preserve a structure* S *with respect to relata* $b_1, ..., b_m \in A(w)$ if and only if we have $R\big(b_1, ..., b_m\big) \neq R\big(v(b_1), ..., v(b_m)\big)$ if S is a relation R, or $v\big(f(b_1, ..., b_{m-1}\big) \neq f\big(v(b_1), ..., v(b_{m-1})\big)$ if S is a function f.
This negation agrees with the intuition because the definiendum already states part of the condition that follows, namely that $b_1, ..., b_m$ are relata of the structure S in $A(w)$, which implies that $(b_1, ..., b_m) \in R$ if S is a relation and that $f(b_1, ..., b_{m-1})$ exists in $A(w)$ if S is a function, meaning that the structure is satisfied before the variation.

For applications it is again important to see that the definition can fail for two reasons. First, it could be the case that one or more of the $v(b_i)$ do not exist in $A(e')$, if the corresponding aspect disappears. Second, the identities may fail to hold.

Definition 9.A.3. A property $a \in \mathcal{A}$ is a S-property if and only if the following condition holds:
A variation *does not preserve* S with respect to relata $b_1, ..., b_m$ *if and only if* the variation *does not preserve* a relative to $b_1, ..., b_m$.

This condition needs to hold true for all variations and all relata. This means that it needs to hold true for all variations of all worlds w in the class W that instantiate relata of the structure S. Definitions 9.A.1 to 9.A.3 allow us to define Conditions (D1′) and (D2′) in more detail.

Definition 9.A.4. A mathematical structure $\mathbb{S}$ is a mathematical structure of the world if and only if the following two conditions hold:

(D1′) The domains $\mathcal{A}_i$ of $\mathbb{S}$ are subsets of $\mathcal{A}$.

(D2′) For every S_j, there is a S_j-aspect in $\mathcal{A}$.

Here, $\mathcal{A}$ denotes the set of all properties of the worlds in W as defined in (9.6).

9.B. Objections

In this appendix, we provide the Lemmas that underlie the explanations in Section 9.7.

Lemma 9.B.1. If a property a is a S-property, every world that instantiates relata of S needs to instantiate a relative to these relata.

Proof. Let a be a S-property and w be any world that instantiates relata $b_1, ..., b_m$ of S. Definition 9.A.3 holds true for all variations of all worlds that instantiate relata of the structure S. Because w instantiates relata, Definition 9.A.3 applies to any variation that maps from w to any other world. Let v be any such variation. This variation either preserves S with respect to relata $b_1, ..., b_m$, or it does not preserve S with respect to relata $b_1, ..., b_m$.

Because a is a S-property, if v preserves S with respect to relata $b_1, ..., b_m$, then it preserves a relative to $b_1, ..., b_m$. But according to Definition 9.A.1, this can only be true if $a \in \mathcal{A}(w)$ relative to $b_1, ..., b_m$ (cf. Footnote 17 for details). If, on the other hand, v does not preserve S with respect to relata $b_1, ..., b_m$, then it does not preserve a relative to $b_1, ..., b_m$. But according to Definition 9.A.1, this too can only be true if $a \in \mathcal{A}(w)$ relative to $b_1, ..., b_m$. Thus both cases imply $a \in \mathcal{A}(w)$ relative to $b_1, ..., b_m$. Thus the result follows. □

The condition that corresponds to the automorphism group of a structure being trivial in the full formal setting of our definition introduced in Appendix 9.A is that no variation of a structure, other than the identity, preserves this structure. For this case, we have the following lemma.

Let S be a structure over a domain $\mathcal{A}_0$, and assume that S contains at least two sets of relata that are properties of least two worlds.

Lemma 9.B.2. If no variation preserves S with respect to any of its relata, no S-property exists.

Proof. Let w_1 and w_2 be worlds that instantiate the relata of S. Lemma 9.B.1 implies that if there is a S-property a, both of these worlds need to instantiate a relative to the relata that they instantiate. Consider now a variation from w_1 to w_2 which maps the relata instantiated in w_1 to the relata instantiated in w_2. According to Definition 9.A.1, this variation preserves a relative to the relata instantiated in w_1 (cf. Footnote 17). Thus there is a variation that preserves a relative to said relata. If a is a S-property, Definition 9.A.3 furthermore implies that the variation preserves S with respect to those relata. This contradicts the antecedent of the claim in the Lemma. Hence no S-property can exist. □

Part IV.

On Artificial Consciousness

10. The Case for Neurons: A No-Go Theorem for Consciousness on a Chip

Johannes Kleiner and Tim Ludwig [1]

The question of whether artificial intelligence (AI) systems are conscious has emerged as one of critical scientific, philosophical, and societal concern. While empirical support to differentiate theories of consciousness is still nascent and while current measures of consciousness (the simplest example of which is interpretation of verbal reports) cannot justifiably be applied to AI systems, our best hope for reliable answers is to link AI's potential for consciousness with fundamental properties of conscious experience that have empirical import or philosophical credibility.

Significant progress in this regard has already been achieved. In (Chalmers, 2023a), David Chalmers assesses evidence for or against AI consciousness based on an extensive array of features that a system or organism might possess or lack, such as self-report, conversational ability, general intelligence, embodiment, world or self-models, recurrent processing, or the presence of a global workspace. In (Wiese, 2024), Wanja Wiese proposes a criterion for distinguishing between conscious and non-conscious AI, anchored in the desiderata of the neuroscientific Free Energy Principle.[2]

[1] Forthcoming in *Neuroscience of Consciousness*.

[2] These are examples of research whose aim is to evaluate *whether* AI systems of the more recent form are or can be conscious. Other interactions between AI research and consciousness science include the use of AI inspired tools and concepts to model consciousness, for example (L. Blum & Blum, 2022; Ji et al., 2024; Juliani, Kanai, & Sasai, 2022), and studies of how models of consciousness might help to build better AI, for example (L. Blum & Blum, 2023; Juliani, Arulkumaran, Sasai, & Kanai, 2022; Mollo & Millière, 2023). The question of whether machines in general can be conscious has guided much of the debate in philosophy of mind over the previous decades, cf. (Block, 1980; Bronfman, Ginsburg, & Jablonka, 2021; Chalmers, 2010; Clancey, 1993; A. Clark, 1998; Dennett, 1991; Haugeland, 1989; Holland, 2003; Penrose, 1989; Searle, 1980; Edelman, 1989; Tegmark, 2017; Turing, 1950), among others.

In this paper, we propose a result of similar nature, which however does not rely on system features and how they relate to consciousness, but on a general property of consciousness: *dynamical relevance*. Here, *dynamical* refers to the temporal evolution (the dynamics) of a system's states as described by a theory of consciousness. Consciousness is *relevant* to a system's time evolution if the time evolution with consciousness differs from the time evolution without consciousness. Whether or not consciousness is dynamically relevant depends on the theory of consciousness under consideration, and in how far this theory implements consciousness-dependent changes of the dynamical evolution, as compared to a reference theory that addresses the same states.

What sets AI systems apart in the context of consciousness is not the specific computational architecture that is employed; architectures that closely resemble the mammalian brain's computational structure can arguably also be used, after all (K. J. Friston et al., 2022). Instead, the distinctive aspect is the hardware on which an AI architecture operates, namely CPUs, GPUs, TPUs, or other processors. This hardware is designed and verified to ensure that the system's dynamics evolve precisely as described by a computational theory during what is known as *functional* and *post-silicon verification*. These verification processes ensure that the design of the chip (the layout of integrated circuits in terms of semiconductors), as well as the actual product (the processing unit after production), yield dynamics exactly as specified by the computational theory. Any dynamical effects that violate the specification of this theory are excluded or dynamically suppressed by error correction.

Our result is an example of a no-go theorem similar to those used in physics. That is, it is a formal theorem that proves a conclusion to hold based on formal assumptions. In our case, these assumptions comprise dynamical relevance of consciousness, as well as formal statements of functional and post-silicon verification.

No-go theorems serve an important role in scientific progress in physics. This role is not necessarily to establish a conclusion beyond doubt, but to direct research and attention to the assumptions that feed into the no-go theorem. Only once such assumptions have been confirmed to hold true, the conclusion of the theorem will be established.[3]

In this spirit, we too do not contend that our result resolves the issue of AI consciousness. Rather, we take our result to point at the theorem's assumptions, most notably dynamical relevance, for further research. If dynamical relevance holds true, then our result does have strong implications. If it does not hold true, our result ceases to apply. In explaining our assumption in Section 10.1, we do give good reasons for why dynamical relevance may plausibly be true, but our explanations are not intended to establish this beyond reasonable doubt. Rather, they are meant to invite further research to establish clarity in respect of this assumption.

Our theorem is mathematical in nature, it rests on formal quantities and a formal proof. And much like formal proofs in other sciences can only be intuitively explained up to a certain point, so can our proof. The following argument is an attempt to explain our proof intuitively, but we would like to stress that this intuition does not capture the res-

[3] We would like to thank Ryota Kanai for introducing the notion of no-go theorems to consciousness science.

ult in full. In fact, the objective of formal modelling is to delineate all concepts involved in intuition carefully, so as to present a theorem that underwrites the intuition both in scope and precision.

(A1) Verification of processing units ensures that any dynamical effects that change the computational dynamics of a processing unit are precluded or suppressed.

(A2) If consciousness is dynamically relevant, and AI systems are conscious, then there are dynamical effects that change the computational dynamics of an AI system.

(A3) AI systems run on processing units.

(C) If consciousness is dynamically relevant, AI systems cannot be conscious.

The conclusion (C) follows because qua (A3) and (A1), verification ensures that any dynamical effects that change the computational dynamics of an AI system are precluded or suppressed. (A2) states that if consciousness is dynamically relevant, and AI systems are conscious, then there are dynamical effects that change the computational dynamics of an AI system. Therefore, if consciousness is dynamically relevant, then AI systems cannot be conscious. The crucial work of the formalisation we introduce below is to make sure this reasoning is also sound if consciousness' dynamical effects apply on a "level below" the computational level.

In a nutshell, this paper shows that if consciousness makes a difference to how a system evolves in time—as it should if consciousness is to have any evolutionary advantage, for example—then any system design which systematically precludes or suppresses diverging dynamical effects systematically precludes or suppresses the system from being conscious.

Before embarking on the formal research that puts the above reasoning on solid ground, we focus on the new concept of dynamical relevance: we explain it in more detail in Section 10.1, and give reasons for why it may, plausibly, be true.

10.1. What is Dynamical Relevance?

Dynamical relevance is a formal condition. It is defined in Section 10.2.2, once formal preliminaries have been introduced in Section 10.2.1. The goal of this section is to explain and illustrate the concept in non-formal terms, so as to make it accessible to a wide audience.

Dynamical relevance is a relational concept. It describes how something, for example a property, relates to the dynamics of a system, as described by a theory. If that "something" is relevant for the dynamics of the system, then we call it *dynamically relevant*. In contrast, if that "something" is not relevant for the dynamics of the system, then we

call it *not dynamically relevant* or *dynamically irrelevant*. Before applying dynamical relevance to consciousness, let us give two examples of how this notion applies to other properties.

Example 1: A moving car

As an intuitive first example, we consider a hypothetical theory for a moving car.[4] The theory predicts, we presume, how the car behaves as forces are applied to it. In particular, it describes which dynamical trajectory the car takes on a parking lot as forces are applied to its steering wheel and its brake and gas pedals for a given initial position and velocity.

How much load we add to the car is not predicted by the moving-car theory, it requires an extension of this theory that is also capable of dealing with load. If one puts a heavy box into the trunk of the car, the car's dynamical trajectory will be different from its dynamical trajectory with an empty trunk. This difference might be small and hard to notice or large and easy to notice; for example, in the case of a Moose test, a heavy box in the trunk could make the difference between tipping over and not tipping over. In any case, as the load of the car makes a difference to the dynamics of the car, the moving-car-plus-load theory introduces a new variable that is dynamically relevant with respect to the moving-car theory.

The colour of the car's seats is also not predicted by the moving-car theory, and if that should be taken into account, an extended model with a new variable that describes said colour is required as well. For example, the seats could be coloured in black, blue, or red. In contrast to the car's load, however, the moving-car-plus-colour theory will not make changes to the dynamical trajectory of the car; the car's dynamical trajectory will be the same for all seat colours. Thus, as the seat colour doesn't make a difference to the dynamics of the car, according to the moving-car-plus-colour theory, the seat colour is dynamically irrelevant with respect to the moving-car theory.

To summarise, for the hypothetical extensions of the moving-car theory outlined above, the car's load is dynamically relevant, whereas the seats' colour is dynamically irrelevant. We emphasise that the specification of the reference theory is important. With respect to a more elaborate moving-car theory that takes into account the driver and their psychology for the prediction of the car's dynamical trajectory, the seats' colour might very well make a difference for the dynamics of the car and, thus, be dynamically relevant.

Example 2: An electrical circuit

As a more scientific example, we consider an electrical circuit. In an electrical circuit, voltages and charge currents are typically described by electrical circuit theory. For example, Ohm's law $V = R \cdot I$ relates the voltage drop V across an electrical resistor with resistance R to the charge current flow I through the resistor. Besides the resistor, the electrical capacitor is another important circuit element. A capacitor stores electrical

[4] We thank Wanja Wiese for suggesting this example when discussing our manuscript.

charge Q, when a voltage V is applied to it; the capacitor's capacitance C determines the amount of charge that is stored for a given voltage $Q = CV$.

Based on the two circuit elements, resistor and capacitor, one can build a simple electrical circuit: a so called RC-circuit, where a capacitor is effectively connected to itself but only via the resistor. When the capacitor is initially charged up to the voltage V_0, it will decay on a timescale $\tau = RC$; explicitly, $V(t) = V_0\, e^{-t/\tau}$. This constitutes a model for the capacitor voltage in an RC-circuit; or, for brevity, RC-circuit model.

This model can be extended to take into account further quantities of interest. For example, the resistance of a resistor R depends on the temperature T of the resistor. Temperature is a concept from thermodynamics but not from circuit theory, so it is not part of the RC-circuit model as described above. But the temperature is relevant for the resistance, and hence it is dynamically relevant for the voltage in an RC-circuit; it changes how the voltage evolves over time. Thus a model that extends the RC-circuit model to take into account temperature posits temperature as dynamically relevant. In contrast, if we extended the RC-circuit model to take into account the resistor's colour coating, the new variable would not be dynamically relevant, because the resistor's colour coating is dynamically irrelevant for the voltage in an RC-circuit; it makes no difference to how the voltage or other quantities in the original model evolve in time.

Dynamical relevance of consciousness

Having clarified the concept of dynamical relevance in general contexts, we can now discuss its application in consciousness science. For brevity, we will use the term 'dynamical relevance' in what follows to abbreviate the term 'dynamical relevance of consciousness'.

Dynamical relevance (of consciousness) describes the relation between a theory of consciousness and a reference theory on which the theory of consciousness is built, for example a neuroscientific theory that describes those brain functions that operate independently of consciousness. In a nutshell, a theory of consciousness posits consciousness as dynamically relevant, if being conscious makes a difference for the time evolution of a system, as compared to what the reference theory, that does not contain consciousness, would prescribe.

A simple example of a theory of consciousness that posits consciousness to be dynamically relevant is a theory which proposes that consciousness is a specific cognitive function that would be absent if systems did not possess consciousness. Another simple example is a theory of consciousness which posits that consciousness is something non-physical and endows consciousness with a causal effect on physical states.

Relation to other Properties

Consciousness can be dynamically relevant in both physicalist and non-physicalist ontologies. That is, it is *ontologically neutral*. By endorsing dynamical relevance one is not

committed to any specific ontology. As we will now show, dynamical relevance is furthermore implied by other (important) concepts in both physicalist and non-physicalist concepts. Therefore, dynamical relevance is a *weaker* assumption than those concepts. It is easier to accept and less demanding than these other concepts.

In *physicalist contexts*, dynamical relevance is implied by at least three concepts. First, it is implied by strong emergence. That is the case, because the "fundamental higher-level causal powers" (O'Connor, 2021b, Sect. 4), which exist in the case of strong emergence, make a difference to the time evolution of the substrate states.

Second, dynamical relevance can also be implied by some forms of weak emergence. It is arguably implied, for example, by the information decomposition approach to causal emergence (P. A. M. Mediano et al., 2022). In this approach, even weak emergence induces downward causation. If downward causation implies that there are causal effects of the higher-level property on the lower-level property, then the higher-level property is dynamically relevant to the lower-level property.

Finally, dynamical relevance is also implied by the assumption that consciousness has intrinsic or functional value (Cleeremans & Tallon-Baudry, 2022), which motivates agents and guides their behaviour. That is the case because an agent's behaviour is part of the agent's dynamical trajectory. Therefore, if "it is only in virtue of the fact that conscious agents 'experience' things and 'care' about those experiences that they are 'motivated' to act in certain ways" (Cleeremans & Tallon-Baudry, 2022, p. 1), then consciousness is dynamically relevant.

In *non-physicalist contexts*, dynamical relevance (of consciousness) is implied by a violation of an ontological assumption known as 'causal closure of the physical' or 'completeness of the physical' (Robb, Heil, & Gibb, 2023). This assumption states that for every physical effect, there are sufficient physical causes.

Dynamical relevance is implied by a violation of the causal closure of the physical, because if the physical is not causally closed in virtue of consciousness, there are physical effects at least one of whose jointly sufficient causes is consciousness—usually conceived of as a property or substance separate from the physical properties or substances in this context. But a cause makes a difference to the time-evolution of its effect. Hence it follows that consciousness makes a difference to the time evolution of some physical effects: the time evolution with consciousness differs from what it would have been without consciousness. Thus, if the physical is not causally closed in virtue of consciousness, consciousness is dynamically relevant.

Is dynamical relevance plausibly true?

Our no-go theorem is predicated on dynamical relevance; it only applies *if* dynamical relevance holds true, and its conclusions apply to AI systems only in this case.

This paper is not intended to establish dynamical relevance as true. A key function of no-go theorems is to point to the underlying assumptions, and this is exactly what we take the main point of our theorem to be.

What we need to do, however, is to give reasons for why it is plausible to assume dy-

namical relevance. Some of these reasons have already been given above. Because dynamical relevance follows from other assumptions that are taken to be valid—because it is a *weaker* assumption—, it is plausibly true. However, there are also more direct reasons for this, which we review in this section.

Consider, as a simple example, an experiment which relies on a subject's reports on her conscious experiences. Let us assume that the subject is shown some stimulus followed by a mask, and that she has to press a button to indicate whether she has consciously perceived the stimulus, or not, across various trials. Throughout the trials, we might measure her EEG signal, so as to carry out an analysis that distinguishes EEG activity in the case of conscious perception from EEG activity in the case of unconscious perception. This analysis might target a theory of consciousness, so as to confirm or refute whether the difference in EEG signal is aliened with the theory's predictions or retrodictions about this case.

A necessary condition for such an analysis to be possible is that the report—the pressing of a button, in this case—*can* depend on whether the subject has consciously perceived a stimulus, or not. Put in terms of the theory of consciousness that a study aims at, we may say: A necessary condition, for the above analysis to be possible, is that the report (or EEG data for that matter) depends on whether the subject is experiencing the stimulus consciously or not (according to the theory, if it were true). If the time-series of reports and EEG data does not depend on consciousness, the experiment cannot have any weight in supporting the theory. In other words, the theory must posit consciousness as relevant to the report or EEG data (or both). And because report and EEG data are part of the dynamics of theories from natural sciences, the theory of consciousness must posit consciousness to be dynamically relevant. Dynamical relevance is likely a precondition for the experiment and the analysis to work as intended. Further details are needed to cash out this example, and to see if it indeed applies. But we take it to show that dynamical relevance is at least plausibly true.

More generally, we may say that any empirical investigation of consciousness relies on *measures of consciousness* (Irvine, 2013) to infer the state of consciousness of a subject (some information about the subject's conscious experience, that is). An experiment may use objective measures of consciousness that rely on behavioural or neural markers, or subjective measures of consciousness that rely on a subject's reports about their conscious experience. Both types of measures rely on data that is part of the dynamics of the physical. And for a measure of consciousness to work as expected—to allow us to infer something about the state of consciousness of a subject—, consciousness must make a difference to the data that feeds into the measure. It must make a difference to the dynamics that explain the data, and hence be dynamically relevant, with respect to a theory that contains such explanation.

The same argument can be made not only for scientific investigations, but for any kind of intersubjective exploration of conscious experiences. Debating consciousness relies on certain dynamics of the vocal cord (among many other things), making art about consciousness makes use of behaviour. All of these cases are part of the dynamics of an organism, and if the dynamics are to depend on consciousness, consciousness

needs to be dynamically relevant.[5]

The upshot of these arguments is that dynamical relevance could well be a *necessary condition* for the type of activities we carry out when engaging in *empirical* scientific studies of consciousness. These arguments do not show that dynamical relevance is true. For all we know, there is the possibility that it isn't. But if it isn't, the empirical investigation of consciousness—and with it the science of consciousness—might not make sense; a necessary condition for its possibility would likely be violated.

Current Theories

The above arguments do not depend on any specific theory of consciousness. But it is interesting to ask what current theories of consciousness say about dynamical relevance.

First, it is important to note that empirical tests of theories of consciousness presume that consciousness is dynamically relevant according to these theories. That is the case, because they assume that whatever is measured can corroborate or falsify a theory, or speak in favour of one theory rather than another. For this to be possible, consciousness must make a difference to the data. And because the data is drawn from the physical dynamics of a system, consciousness must be dynamically relevant.

Second, we can consider the metaphysics of theories of consciousness. In cases where these are clear, they do, in our eyes, imply dynamical relevance. Consider, as an example, Integrated Information Theory (IIT) (Oizumi et al., 2014). IIT assumes that experience is primary and physics—or better, physical descriptions—are secondary. In a sense, only experience exists, in the form of cause-effect-structures. Hence it should be the case that experience makes a difference to the physical dynamics, so that conscious experience is dynamically relevant.

Another example is Global Neuronal Workspace Theory (GNW) (Dehaene et al., 2011). Here, too, we think, the metaphysical interpretation implies dynamical relevance. GNW assumes that conscious experiences are tied to a global neuronal workspace, "consisting of a distributed set of (...) neurons characterised by their ability to receive from and send back to homologous neurons in other (...) areas horizontal projections through long-range excitatory axons" (Dehaene et al., 2011, p. 56). Organisms that posses a workspace are conscious, while organisms that do not posses a workspace are not conscious, according to the theory. Hence whether or not a system is conscious makes a difference to a system's information processing architecture and, a fortiori, to the system's dynamics.

The only thing which speaks against dynamical relevance among current theories of consciousness, in our eyes, is their mathematical formulation (in those very limited cases where a mathematical formulation exists).

[5] This argument can be strengthened by considering what is required to distinguish two or more theories of consciousness empirically, cf. (Kleiner & Hartmann, 2023), where however dynamical relevance is referred to as 'empirical version of the closure of the physical' in (Kleiner & Hartmann, 2023), and formulated in more generality than we do here.

Consider, for example, Integrated Information Theory (IIT). The mathematics of IIT is given in terms of an unwieldy algorithm that takes as an input a physical description of a system *as given by some reference theory*, and provides as output a mathematical description of the conscious experience of that system. An analysis of the mathematics that underlie this algorithm shows that the algorithm defines a map which goes from the description to the descriptions of conscious experience (Kleiner & Tull, 2021; Tull & Kleiner, 2021).

Therefore, according to IIT's mathematics, consciousness is not dynamically relevant. The physical evolution of the systems are exactly as they are in the reference theory that provides the input to IIT. No change whatsoever is introduced to these dynamics by the theory. The mathematics of IIT do not instantiate dynamical relevance.

In our view, this is an issue of the mathematical formulation that IIT applies. The mathematics do not naturally align with the metaphysical foundation of the theory, and the exact same formal properties which speak against dynamical relevance are the source of other issues, most notably issues with falsifying the theory, cf. (Kleiner & Hoel, 2021), and issues related to the unfolding argument, more generally (Doerig et al., 2019). The mathematics of IIT may need to be revised, at the very least to instantiate dynamical relevance, so as to resolve these problems with falsification.

Definitions

We conclude this section with a pointer to the places in the manuscript where the precise definition of dynamical relevance is given: in Section 10.2.2, Definitions 10.2.1 and 10.2.2. Definition 10.2.1 is epistemic. It defines the concept of dynamical relevance with respect to a theory of consciousness, relative to some underlying neuroscientific theory, independently of whether either of the theories is true. Definition 10.2.2 then builds on this epistemic definition to provide an ontic definition. This definition is about whether consciousness is actually dynamically relevant. What is crucial in Definition 10.2.2 is that it suffices that there is *some* reference theory with respect to which the true theory of consciousness satisfies Definition 10.2.1. This is sufficient to prove our result, Theorem 10.2.4. Referencing the actual world is important in the context of this result because post-silicon verification is about what actually happens, once a processing unit has been manufactured.

10.2. No-Go Theorem

10.2.1. Formal Preliminaries

The central notion which underlies our result is that of the time evolution of a system's states. Given a scientific theory T and a system S within the scope of the theory, we denote by $k_T(S, s)$ the *dynamical evolution* (also called 'trajectory') of S with initial state s. This dynamical evolution describes how the state s evolves in time according to T. An example is the evolution of a brain state according to a neuroscientific theory. We will

abbreviate $k_T(S, s)$ by k_T if it is clear from context that we're talking about one system and one initial state.

The class of scientific theories which is relevant in the present context are theories of consciousness, on the one hand, and neuroscientific theories on which theories of consciousness are built, on the other hand. These neuroscientific theories are theories that a theory of consciousness makes use of to explain how consciousness relates to the brain, and to which it refers for all explanations that do not involve consciousness: the theories that have been developed in neuroscience or other natural sciences. We use the symbol Υ to denote all such theories that are relevant for AI or consciousness, and refer to the theories in this class as *reference theories*, because they are the theories that a theory of consciousness can refer to. Examples are theories of neuroscience, biology, chemistry, computer science and physics.

Different theories describe systems at different levels (List, 2019), and in some cases, the states of a system posited by one theory T (the "lower" level) can (in principle) be mapped to states of another theory T' (the "higher" level). If this is the case, we write $T < T'$. Because dynamical evolutions are sequences of states, if $T < T'$, we can map any dynamical evolution k_T of T to a (not necessarily dynamical) evolution of T', which we denote as $k_T|_{T'}$.

We assume that there is a reference theory $T_F \in \Upsilon$ that can be mapped to states of any other reference theory in Υ, which means that $T_F < T$ for all $T \in \Upsilon$. For lack of a better term, we will refer to this theory as a *fundamental reference theory*, but emphasise, that it does not have to be "the true" fundamental theory. The requirement that $T_F < T$ for all $T \in \Upsilon$ is only an epistemic requirement that expresses relationships been theories in Υ, and leaves open whether T_F, or any other theory in Υ for that matter, is the true theory which correctly describes the actual dynamics. Whether or not this can be the case depends precisely on the question of whether consciousness is dynamically relevant. What justifies the assumption that there is a theory whose states can be mapped to states of the other theories (whose states *ground* the states of all other theories, one might say) is that the states of quantum theory can, in principle, be mapped to states of all physical theories in Υ. That is because quantum theory is what underlies condensed-matter theories as far as they are relevant for semi-conductors and integrated-circuit design of processors. So, for all practical purposes, we can think of T_F as quantum theory. We remark that the requirement of a relationship of states is much weaker than any reductive assumption.

Finally, we assume that there is a fact to the matter of what the real (that is: actual) dynamics of any system are, even if that fact may not be knowable. We denote the description of the real dynamics in terms of the states of any reference theory $T \in \Upsilon$ (any "level" of description, so to speak) by $k^*|_T$. If $T < T'$, the description of the real dynamics in terms of the states of both theories are compatible, that is $k^*|_T|_{T'} = k^*|_{T'}$.

10.2.2. Dynamical Relevance

Theories of consciousness (tocs), sometimes also called models of consciousness, express a relation between a description of a system, on the one hand, and a description of its conscious experiences, on the other hand. The latter could be a description of its phenomenal character (cf. e.g. (Kleiner & Ludwig, 2024; Lee, 2021)), or simply an expression of whether a system S has conscious experiences at all. Together, both descriptions constitute a state s of the toc. Because a toc expresses a relation between a description of a system and a description of its conscious experiences, the state s contains both a non-experiential and an experiential part, which we refer to as *reference state* and *state of consciousness*, to have a simple terminology that is free of metaphysical burden. The dynamical evolution $k_M(S, s)$ of a system S in a state s of the theory/model of consciousness M expresses how the reference state and the state of consciousness relate according to the theory.

Because tocs contain a reference description of a system at some level, for every toc M, there is at least one reference theory $T_R \in \Upsilon$ such that the physical part of any state s of M, and therefore also any dynamical evolution k_M, can be expressed in T_R. We denote this state by $s|_{T_R}$ and the expression of the reference part of the trajectory k_M in terms of T_R by $k_M|_{T_R}$. So, $k_M|_{T_R}$ is what M says about the evolution of reference states on T_R's level of description. We call any such T_R an *underlying* reference theory of M.

To offer an alternative perspective that might be helpful to illustrate this notation, consider again that any theory of consciousness M expresses a relation between a description of a system and a description of its conscious experience, or if framed in the terminology we have just introduced: a relation between a reference state and a state of consciousness. Let us suppose that the former constitute a set $\tilde{P}$ and that the latter constitute a set E. Here we are adding a '$\sim$' on top of P because the states which the theory of consciousness uses might not be identical to the states that any reference theory uses; there could be simplifications, for example. What needs to be the case, however, is that these states can be mapped to the states of *some* reference theory T_R. The states of the reference theory are what the theory of consciousness "means" when addressing reference states, so to speak. Let us assume that the reference states of T_R form a set. A trajectory k_M of M is a trajectory over $\tilde{P} \times E$. By restricting to $\tilde{P}$ and then mapping to P, we obtain a trajectory over P. This is what the symbol $k_M|_{T_R}$ denotes: it is what the trajectory of M implies for the time evolution as expressed in terms of the states of the reference theory T_R.[6]

Independently of what the description is that a toc applies on the side of consciousness, there is a fact to the matter of whether a system is conscious or not when in a trajectory $k_M(S, s)$. This means: whether the system S has conscious experiences at least at one point of time in the dynamical evolution $k_M(S, s)$. Making use of the important link between tocs and reference descriptions, we can say that a system S *is conscious* in a dynamical trajectory k_{T_R} of the reference theory iff there is a dynamical

[6] We are grateful to an anonymous reviewer for pointing out that this perspective might be helpful to include.

evolution k_M of M such that (a) we have $k_M|_{T_R} = k_{T_R}$ and (b) the system is conscious in k_M.

Whether a toc has anything original to say about the dynamical evolution of its reference states, or simply presumes the dynamical evolution of a reference theory—of an underlying neuroscientific theory, that is, in most cases—, is precisely the question of dynamical relevance, defined as follows. Let M denote a toc and $T_R \in \Upsilon$ a reference theory thereof.

Definition 10.2.1. *Consciousness is dynamically relevant according to M with respect to T_R* iff

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T_R} \neq k_{T_R}\,.$$

Here, the right-hand-side is short-hand for $k_M(S,s)|_{T_R} \neq k_{T_R}(S, s|_{T_R})$, where $s|_{T_R}$ denotes the restriction of the state s of M to T_R. The left-hand side is a shorthand for 'S is conscious in $k_M(S,s)$', meaning that there is at least one point of time in $k_M(S,s)$ so that S has a conscious experience at that time according to M. The condition has to hold for all dynamical trajectories k_M of M, meaning for the dynamical trajectories of all systems S in the scope of M, and all states s of these systems.

This definition expresses the intuition that if S is conscious according to a toc M, then the dynamical evolution as specified by M differs from the dynamical evolution as specified by the underlying neuroscientific theory alone.

We have already referenced the 'real' dynamics of a system and introduced the symbol $k^*|_{T_R}$ to denote what the real dynamics of a system would look like in terms of the states of T_R. There is also a fact to the matter of whether a system in a trajectory k^* is conscious and how conscious experiences relate to the physical. That is, there is a 'true' or 'real' theory of consciousness, which we denote by M^*. As in the physical case, M^* may be unknown or unknowable. We will denote its dynamical evolutions by k_{M^*}. Because these describe what really happens, we have $k_{M^*}|_{T_R} = k^*|_{T_R}$ for all T_R. Using M^*, we can define dynamical relevance simpliciter:

Definition 10.2.2. *Consciousness is dynamically relevant* (CDR) only if it is dynamically relevant according to the 'true' toc M^* with respect to *some* reference theory $T_R \in \Upsilon$.

10.2.3. Functional and Post-Silicon Verification

What is unique about AI systems in the present context is not the particular architecture that is employed; AI can also be built on architecture derived from the brain; cf. e.g (K. J. Friston et al., 2022). What is unique is rather that the architecture runs on CPUs, GPUs, TPUs or other processors that have been designed and *verified* in the lab.

There are two major verification steps in processor development, called functional and post-silicon verification. *Functional verification* (Mishra & Dutt, 2005; Wile, Goss, & Roesner, 2005) is applied once the design of a processor in terms of integrated circuits has been laid out, but before the manufacturing phase begins. It applies simulation

tools, formal verification tools and hardware emulation tools to ensure that the design of the chip meets the intended specifications as described by a computational theory T_{comp}. *Post-silicon verification* (Mishra, Morad, Ziv, & Ray, 2017; Mitra, Seshia, & Nicolici, 2010) is applied after the silicon waver has been fabricated. It applies in-circuit testing, functional testers, failure analysis tools and reliability testing, among other things, to ensure that the physical product works as T_{comp} would have it. The theory T_{comp} is specified by a processing unit's instruction set architecture, present-day examples of which are the ARM instruction set architectures on which on most data centre servers run, or the X-86 instruction set architecture on which most desktop devices run.

Functional verification is a theoretical endeavour. It applies simulation and emulation tools based on a theoretical account on how the substrate, on which a processor is to be built, behaves. Because this substrate is a semi-conductor, this theoretical account is based on quantum theory. Put in terms of dynamics, functional verification aims to ensure that whatever happens in the quantum realm, or below, implements or is compatible with the dynamics as described by T_{comp}, formally:

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \tag{10.1}$$

for all dynamical evolutions of a processor S. This condition could fail, for example, because of leakage currents, most notably those created by tunnelling of electrons through a transistor's gate oxide layer. Tunnelling is an effect described by quantum theory, and needs to be controlled for in order to ensure transistors implement a choice of T_{comp}.

Post-silicon verification, on the other hand, is applied to a chip once it has been built. It ensures that the dynamics of the actual physical product comply with T_{comp}. Making use of the k^* notation to denote the actual dynamical evolution of a system, post-silicon verification enforces that

$$k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \tag{10.2}$$

for all dynamical evolutions of a processor S.

Being an AI system means running on CPUs, GPUs, TPUs or other processors that have been designed and verified. That's what makes the system "artificial". And because processor dynamics compose (the output of one is the input of the next), verification holds for AI systems as well: there is an underlying computational theory T_{comp} that accounts for what "happens" on the processors while the system is running, and the computational dynamics satisfy (10.1) and (10.2).

10.2.4. AI Consciousness

With all this in place, we can formulate the question that is being asked precisely. The term 'artificial intelligence' is used very broadly, comprising many different computational architectures and applications. What one means when one asks whether an AI system is conscious is whether the computational architecture that is applied by this system, with the specific quirks of its implementation and training, potentially in a specific task, has conscious experiences. The architecture and these specifics determine

the computational dynamics the system is capable of. Thus, the question is whether the system has a computational evolution $k_{T_{\text{comp}}}$ such that it is conscious in this computational evolution according to a theory of consciousness M; cf. Section 10.2.2 for a definition of what this means in terms of dynamics k_M of M.[7] In summary:

Definition 10.2.3. An *AI system S is conscious* according to a theory of consciousness M only if there is at least one dynamical evolution $k_{T_{\text{comp}}}$ in which the system is conscious according to M.

This is a very weak condition, which however has one important consequence: that the question of AI consciousness is determined by facts on the computational level and above; it is independent of what happens on a sub-computational level. That is, if we have a a trajectory k_{T_R} on a sub-computational level ($T_R < T_{\text{comp}}$) with $k_{T_R}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$ then S is conscious in $k_{T_{\text{comp}}}$ only if it is conscious in k_{T_R}.

10.2.5. No-Go Theorem

Our main result is the following theorem.

Theorem 10.2.4. If consciousness is dynamically relevant, then AI systems aren't conscious.

Before giving the proof, we first illustrate the result for the simpler case where consciousness is dynamically relevant with respect to the computational level T_{comp} itself. The power of the theorem is to extend this result to all other cases. Subsequent to this illustration, we prove a lemma needed for the main theorem, and then proceed to prove the theorem itself.

So let us consider the case where T_R in Definition 10.2.2 is T_{comp}. The following chain of reasoning assumes that consciousness is dynamically relevant (Definition 10.2.2) with respect to T_{comp}.

Let S be an AI system. Because of post-silicon verification (10.2), all of the dynamical evolutions of S satisfy

$$k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \,. \tag{10.3}$$

Application of Definition 10.2.2 for the case $T_R = T_{\text{comp}}$ implies, via Definition 10.2.1, that if S is conscious in a k_{M^*}, then $k_{M^*}|_{T_{\text{comp}}} \neq k_{T_{\text{comp}}}$. The converse of this statement is that if $k_{M^*}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$, then S is not conscious in k_{M^*}. From the paragraph before Definition 10.2.2, we have $k_{M^*}|_{T_R} = k^*|_{T_R}$ for all T_R. Setting $T_R = T_{\text{comp}}$, this gives $k_{M^*}|_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}}$, which is why the identity (10.3) establishes the prerequisite of the above condition for all dynamical evolutions of S. Therefore, it follows that S is not conscious in any k_{M^*}. Thus, Definition 10.2.3 implies that S is not conscious, as claimed.

The remainder of this section is devoted to the proof of the theorem in the general case. To this end, we first state and prove the following lemma.

[7] The point here is to restrict downwards, not upwards. Any question "above" the computational level can be posed in terms of computational dynamics.

Lemma 10.2.5. *Dynamical relevance passes downward*, in the sense that if $T_R < T'_R$ and consciousness is dynamically relevant according to M with respect to T'_R, then it is also dynamically relevant according to M with respect to T_R.

Proof of the Lemma. Consciousness is dynamically relevant according to M with respect to T'_R, iff

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T'_R} \neq k_{T'_R} \,.$$

Because $T_R < T'_R$, there is a function which maps states—and therefore also dynamical evolutions—from T_R onto T'_R. Therefore, we have

$$k_M|_{T'_R} \neq k_{T'_R} \;\Rightarrow\; k_M|_{T_R} \neq k_{T_R} \,.$$

Together with the above, this gives

$$S \text{ is conscious in } k_M \;\Rightarrow\; k_M|_{T_R} \neq k_{T_R} \,,$$

which is the case iff consciousness is dynamically relevant according to M with respect to T_R. □

We now proceed to the proof of the theorem.

Proof of the Theorem. We first consider the case where T_R in Definition 10.2.2 is T_F.

Let S be an AI system. Because of functional and post-silicon verification, we have

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}} \tag{10.4}$$

for all dynamical evolutions of S. Because consciousness is (by assumption) dynamically relevant and we have assumed $T_R = T_F$, Definition 10.2.1 applies to give

$$S \text{ is conscious in } k_{M^*} \;\Rightarrow\; k_{M^*}|_{T_F} \neq k_{T_F} \tag{10.5}$$

for all dynamical trajectories k_{M^*} of M^*.

Let us now assume that S is conscious in some trajectory k_{M^*} of M^*. According to the last implication, we thus have

$$k_{M^*}|_{T_F} \neq k_{T_F} \,.$$

Because $T_F < T_{\text{comp}}$, we can map both of these trajectories to T_{comp}. For $k_{M^*}|_{T_F}$, this gives

$$\begin{aligned} k_{M^*}|_{T_F}|_{T_{\text{comp}}} &= k^*|_{T_F}|_{T_{\text{comp}}} \\ &= k^*|_{T_{\text{comp}}} = k_{M^*}|_{T_{\text{comp}}} \,, \end{aligned}$$

where we have made use of identities established in Sections 10.2.1 and 10.2.2. Equation (10.4) furthermore establishes that

$$k_{M^*}|_{T_{\text{comp}}} = k^*|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \,.$$

The two facts that (a) $k_{M^*}|_{T_{\text{comp}}} = k_{T_{\text{comp}}}$ and (b) that S is conscious in k_{M^*} establish that S is conscious in $k_{T_{\text{comp}}}$.

Equation (10.4) also establishes that

$$k_{T_F}|_{T_{\text{comp}}} = k_{T_{\text{comp}}} \, .$$

Because of this equation and $T_F < T_{\text{comp}}$, the implication of Definition 10.2.3 explained in the last paragraph of Section 10.2.4 applies and establishes that S is conscious in k_{T_F}.

Unwrapping what 'S is conscious in k_{T_F}' means by definition, we find that there must be a dynamical evolution $\tilde{k}_{M^*}$ of M^* such that

$$\begin{aligned} &\text{(a) } \tilde{k}_{M^*}|_{T_F} = k_{T_F} \quad \text{and} \\ &\text{(b) } S \text{ is conscious in } \tilde{k}_{M^*} \, . \end{aligned}$$

Together, these two conditions violate (10.5). Thus we have arrived at a contradiction.

The assumptions that went into the derivation of this contradiction were that consciousness is dynamically relevant with respect to the T_F level, that S is an AI system, and that S is conscious in a trajectory k_{M^*} of M. The first assumption is stated as a condition in the theorem. Thus it follows that the latter two cannot be both the case.

Because k_{M^*} was arbitrary, it follows that an AI system S cannot be conscious in any trajectory k_{M^*} of M^*. Consequently, applying Definition 10.2.3, it cannot be conscious at all. This establishes the claim that if consciousness is dynamically relevant with respect to T_F, then AI systems aren't conscious.

It remains to consider all other cases of T_R in Definition 10.2.2. Therefore, let us assume that consciousness is dynamically relevant with respect to some $T_R \neq T_F$. Because $T_F < T_R$ for all $T_R \in \Upsilon$, and because dynamical relevance passes downward (Lemma 10.2.5), it follows that consciousness is also dynamically relevant with respect to T_F. Hence the previous case applies and the result follows in full generality. □

10.3. Objections

In this section, we discuss a few immediate responses to our result.

10.3.1. Verification is imperfect

Verification is an industrial process that may not be perfect: despite functional and post-silicon verification, the actual dynamics of a processor may not adhere to the computational theory targeted by verification in all cases. Verification may leave a bit of wiggle-room for the dynamics to diverge from the computational theory. Could this wiggle-room suffice for consciousness to unfold its dynamical effects?

Any answer to this question depends on how exactly consciousness is dynamically relevant and which imperfections arise in day-to-day verification. It is natural to expect

that consciousness' dynamical relevance is systematic in nature: dynamical effects should systematically occur if a system is conscious and make a systematic difference to how the system evolves in time. The imperfections in day-to-day verification, on the other hand, are likely to be mostly random in nature, meaning that the deviation in dynamical evolution they fail to suppress are random too, both in time (when such a deviation can occur) and in the extent to which they can make a difference. If this is true, it is unlikely that the wiggle-room left open due to imperfections suffices for consciousness to unfold its dynamical effects.

10.3.2. Determinism

One objection to our result takes our result to show or imply that a deterministic system cannot be conscious, and argues that this is very unlikely to be true. Hence the result must be wrong or rest on very weak assumptions, so the objection goes.

This objection fails because our result does not show or imply that deterministic systems cannot be conscious. What prevents a system from being conscious, according to our result, is that its design forces it to comply to a formal system that is independent of consciousness. The system is "locked into" a formal system, so to speak. It cannot deviate from it. Reality is forced to adhere to a theoretical construct, by design.

Our result is fully compatible with deterministic systems, and also with a deterministic relevance of consciousness to a system's dynamics.

10.3.3. Probabilistic processing

Verification as applied in industry targets deterministic computational theories. Would our result also hold in case of verified probabilistic processing?

The mathematical framework we apply is compatible with probabilistic processing: we do not make an assumption as to whether the notions of state and dynamical evolution are deterministic or not; a state may well be a probability distribution and its dynamical evolution a stochastic process. Verification, in this case, implies that a system conforms to the stochastic process as described by a stochastic computational theory. This leaves room for consciousness to have a dynamical effect, but only if this effect conforms to the probability distributions as described by the stochastic computational theory. That is, consciousness may determine how the probability distributions of the stochastic computational theory are sampled, but it cannot change them. As in the case of imperfect verification, we remain sceptical as to whether this limited freedom is compatible with the systematic nature of consciousness' dynamical effects that are to be expected.

10.3.4. Quantum computing

Does our result also hold true in the case of quantum computing? Quantum computing is a young industry and it is not yet clear which type of verification, if any, will need to be de-

ployed. It is likely, however, that any type of verification will need to presuppose a notion of *measurement*, which is an inherently vague concept in quantum theory (Bell, 1990) that is partially external to the account of quantum dynamics by the Schrödinger equation. If consciousness were related to measurement (for example via consciousness-induced dynamical collapse as proposed in (Chalmers & McQueen, 2022)), then verification might leave enough room for consciousness to have a systematic and meaningful effect. If, on the other hand, consciousness is not related to measurement in quantum theory, it is likely that verification of quantum computers to adhere to quantum dynamics will preclude any potential dynamical effects of consciousness; just as in the classical case.

10.4. Conclusion

This paper addresses the question of whether AI systems are conscious. Its objective is to introduce a new formal tool, in the form of a no-go theorem, that may provide an answer to this question which is independent of the specific computational architecture that an AI system utilises, and which does not rely on any specific cognitive feature that an AI system might possess or lack that may be related to conscious experience.

The no-go theorem is based on what we take to be the only property that distinguishes AI systems from other cognitive systems, a property that might well embody the actual meaning of the word 'artificial' in 'artificial intelligence': that the system runs on a substrate that has been designed and verified, rather than naturally evolved.

Ultimately, we believe that any scientific statement about whether a system is conscious needs to be based on a theory of consciousness that is supported by theoretical, philosophical, and most importantly empirical evidence. Consciousness Science[8] searches for such theories. The crucial premise in our result—dynamical relevance—is a *property* which theories ascribe to consciousness, so that our theorem can be regarded as establishing a fact about AI's capability for consciousness for a whole class of theories of consciousness: all those that posit consciousness to be dynamically relevant. Results of this form are important as long as evidence in favour of any single theory of consciousness, as well as evidence to distinguish among them, is still in its early stages, and while the space of possible theories remains only partially explored.

Our result has a few interesting, slightly funny, and potentially relevant implications for AI engineering and AI interpretability. The most notable of these is that our result shows that *if an AI system states that it is conscious, then this cannot be because it is conscious*. That is to say, even if an AI system were conscious, the cause of any such statement cannot be that the AI system is conscious. This follows because if such a cause existed, consciousness would have to be dynamically relevant, in which case our theorem implies that the system isn't conscious. Another implication is that if consciousness has functions that could improve a system's information processing,

[8] Also called *Scientific Study of Consciousness* to emphasise the importance of contributions from humanities, most notably philosophy.

then, to make use of those functions, theories of consciousness should be taken into account when designing the substrate on which an AI system will run.

The question of whether AI systems are conscious is of major societal concern (Association for Mathematical Consciousness Science, 2023). It has important ethical (Bostrom & Yudkowsky, 2018; Metzinger, 2021), legal (Benzmüller & Lomfeld, 2020; Susskind, 2019), and technological consequences, and will likely play a major role in shaping governance of AI and how individuals interact with this technology. Our result aims to deliver a rigorous and justified answer to this question that does not rely on contingent assumptions, such as the truth of a particular theory of consciousness, or the validity of a particular test of consciousness when applied to AI systems. The result relies on the truth of its main assumption, dynamical relevance, further investigation of which is an objective of future research.

11. Consciousness qua Mortal Computation

Johannes Kleiner

11.1. Introduction

A fundamental tenet of general purpose digital computing is that software is separated from hardware, so that the same program or algorithm can be run on any suitable system. This tenet is about to be broken. Contemporary developments in Artificial Intelligence (AI) and AI chip production have led to the identification of a novel concept of general purpose computing, called *mortal computation* (Hinton, 2022). This concept draws a line between the type of computations that contemporary processing units do, and the type of computations that brains and other biological organisms carry out.

Computational functionalism, first defined by Putnam (1967), posits, in a nutshell, that consciousness is a computation. This view has gained popularity again in light of the staggering achievements in AI development in recent years. AI models are computations, so if computational functionalism is true, AI models can—and, depending on the nature of the computation that consciousness is, will—become conscious (Butlin et al., 2023).

Here we show that computational functionalism is not indifferent with respect to the type of computation that consciousness is. We show that if there is any organism that is capable of conscious experiences, but which cannot be programmed—for example, non-human animals; cf. Assumption 11.4.1—, then computational functionalism implies that consciousness is a mortal computation. To establish this result, we make use of a differential definition of mortal computation, as well as general facts about the relation between programs, Turing computation and immortal computation.

Our result challenges the usual understanding of computational functionalism, which

is centered around Turing-like models of computation. If our result holds true, consciousness cannot, according to computational functionalism, be a Turing computation or programmed. Yet, contemporary AI systems and programs are Turing computations. Therefore, this result speaks against the possibility of AI consciousness (though it does not aim to settle the issue due to questions of realization, cf. Section 11.7).

The underlying perspective of this paper is that the discovery of mortal computation by Hinton (2022) may well be a first step towards understanding of a whole new paradigm of computation, potentially as consequential as the Turing-Church-Gödel-Herbrand paradigm of computation of the past nine decades (Gödel, 1934; Church, 1936; Turing, 1937b).[1]

11.2. Mortal Computation

The notion of mortal computation was identified and coined by Hinton (2022, Sec. 9), who describes a learning task that makes use of unknown properties of hardware that vary across systems, such as variations in the connectivity of a system, or variations in non-linear processes in a system. As a result, the parameter values that define the learned computation "are only useful for that specific hardware instance, so the computation they perform is mortal: it dies with the hardware" (Hinton, 2022, p. 13). The general computing paradigm of the past nine decades, in contrast, implies that a computation is largely independent of the hardware on which it is run: "[T]he same program or the same set of weights can be run on a different physical copy of the hardware. This makes the knowledge contained in the program or the weights immortal: The knowledge does not die when the hardware dies" (Hinton, 2022, p. 13).

There is, at this early stage, no constructive definition of mortal computation,[2] but

[1] The question of whether a computation is a Turing *computation* is different from questions regarding Turing *computability*. The former concerns the nature of computations. For example, the question of whether neural computations are Turing computations (Piccinini, 2020). The latter concerns functions, in the mathematical sense of the term, that map natural numbers to natural numbers, and asks whether their value can be computed by a Turing machine. A function is Turing-computable iff there is a Turing computation (meaning: an abstract mathematical model of a Turing machine) that halts on all numbers for which the function is defined, and does not halt when provided with numbers for which the function is not defined. This is the case iff the function is λ-computable (Church, 1936; Turing, 1937a) or general recursive (Gödel, 1934; Kleene, 1936). The definition of Turing-computability of functions leaves open what the computation is that implements the function, which is what this paper is concerned with.

[2] There are two ways of reading Hinton (2022, Sec. 9). On a deflationary reading, a mortal computation is simply a Turing computation that is not known in its entirety to an outside programmer. Call this the epistemic reading of mortal computation. It is suggested by Hinton's emphasis of "large and unknown variations in the connectivity" (ibid.). On a different reading, a mortal computation is a computation that fundamentally transcends some of the constraints of Turing computation, for example the existence of an immutable tape for purposes other than read and write actions, or the existence of a transition function that is Markov, as suggested by Hinton's emphasis on "non-linearities of different instances of hardware" (ibid.). Call this the ontic reading of mortal computation. On the ontic reading, the state of affairs of the hardware is partially unknown to the computation itself. The computation may have to deal with, and make use of, non-Turing properties of the hardware. Both interpretations are compatible with (11.1).

we may consider a differential definition, that helps us distinguish mortal computations in virtue of what they are not. To provide such definition, denote by $\mathcal{C}$ the class of all computations. $\mathcal{C}$ comprises all Turing computations, which we will denote by $\mathcal{C}_{\mathrm{TM}}$ in what follows, as well as other notions of computation, for example, non-deterministic Turing computations, neural computations, analogue computations and the yet-to-be-understood mortal computations.

The core intuition behind immortal computation is "that the software should be separable from the hardware" (Hinton, 2022, p. 13). In practice—in central processing units (CPUs), graphics processing units (GPUs), tensor processing units (TPUs), or data processing units (DPUs)—this separation is enabled by a processing unit's *Instruction Set Architecture* (ISA). An ISA contains specifications of all computations that the processing unit can carry out, and it is with respect to these specifications that programs, operating systems and compilers are defined. To run a program is to run machine code that specifies which of the ISA's computations are to be carried out in which order (call this concatenation) and how the results of computations are to be used by other computations (call this combination). Differences among processing units' performance, design, size, etc., are differences in an ISA's *implementation*. The ISA exists to ensure binary-code compatibility of software despite such differences; it is the boundary between software and hardware.

The computations defined by an ISA constitute a reference relative to which software is defined, and which a class of hardware implements. It ensures that a program can run on different physical copies of the same type of hardware. Computation is immortal precisely because it is defined with respect to such reference. We can formalize this requirement as follows.

Definition 11.2.1. A computation $c \in \mathcal{C}$ is *immortal* iff there is a class of reference computations $c_{\mathrm{ref}} \subset \mathcal{C}$ such that c is a concatenation and combination of these reference computations. A computation c is *mortal* iff it is not immortal.

We will denote the class of immortal computations by $\mathcal{C}_{\mathrm{Imm}}$. Immortal computations are meant to be a subclass of Turing computations, so that we have

$$\mathcal{C}_{\mathrm{Imm}} \subset \mathcal{C}_{\mathrm{TM}} \,. \tag{11.1}$$

Because an immortal computation c is a concatenation and combination of reference computations, every system that can realize an immortal computation c must be able to realize its reference computations c_{ref}. This is the only implication of Definition 11.2.1 we will make use of in what follows. To explicate this implication formally, we denote by *Sys* the class of all systems. This class includes, for example, all CPUs, GPUs, TPUs, and DPUs in use today, as well as all biological organisms. Furthermore, we denote by $\mathcal{C}(S)$ all computations that a system $S \in$ *Sys* can realize or implement. Using such formalism is of advantage because it can be applied to any account of implementation of a computation (Piccinini, 2015). The essential implication of the previous definition then reads as follows.

Implication 1. If $c \in \mathcal{C}$ is immortal, then there is a class $c_{\text{ref}} \subset \mathcal{C}$ such that for all $S \in$ *Sys*

$$c \in \mathcal{C}(S) \Rightarrow c_{\text{ref}} \subset \mathcal{C}(S) \,. \tag{11.2}$$

An important class of immortal computations are computations specified by writing a program in some programming language; computations that are coded, that is, and compiled to run on CPUs, GPUs, TPUs or DPUs. We will simply refer to these computations as 'programs'. Programs are immortal because they are defined with respect to some programming language that in turn is defined, via its compiler, with respect to one or more ISAs.

We denote the class of computations that can be coded with any of the existing programming languages by $\mathcal{P}rog$. Because programs are immortal, we have

$$\mathcal{P}rog \subset \mathcal{C}_{\text{Imm}} \,. \tag{11.3}$$

11.3. Computational Functionalism

Computational functionalism was introduced by Putnam (1967) as the following set of assumptions.

1. "All organisms capable of feeling pain are Probabilistic Automata.
2. Every organism capable of feeling pain possesses at least one Description of a certain kind (i.e., being capable of feeling pain *is* possessing an appropriate kind of Functional Organization).
3. No organism capable of feeling pain possesses a decomposition into parts which separately possess Descriptions of the kind referred to in 2.
4. For every Description of the kind referred to in 2, there exists a subset of the sensory inputs such that an organism with that Description is in pain when and only when some of its sensory inputs are in that subset." (Putnam, 1967, 1975, p. 434)

In giving this definition, Putnam equates Probabilistic Automata with *descriptions* of a system; "[t]he Machine Table mentioned in the Description will then be called the Functional Organization of [a system] S relative to that Description" (ibid.).

The understanding of computation has evolved substantially since Putnam (1967), cf. e.g. (Piccinini, 2015). To connect Putnam's definition to computation as presently understood, and to do justice of it being a definition of computational functionalism, we reformulate Condition 2 in abstract terms, making use of the set $\mathcal{C}(S)$ of computations that a system can realize, which we have introduced above. It is clear from the context of Putnam's definition that it is to apply to all systems, not just organisms in a narrow sense. Denoting the experience of "feeling pain" by e, and the class of systems capable of having this experience by Sys_e, we may hence read Putnam's Condition 2 as follows.

Implication 2. Computational functionalism implies that there is at least one computation $c^* \in \mathcal{C}$ such that, for all $S \in \mathit{Sys}$,

$$S \in \mathit{Sys}_e \;\Rightarrow\; c^* \in \mathcal{C}(S)\,.$$

In Putnam's terms, being capable of realizing c^* is being capable of experiencing pain. Modulo details of sensory input referred to in Putnam's Condition 4, we may say that experiencing pain *is* realizing a computation c^*, or, in more simple terms yet, that the experience of pain is c^*. Nothing hinges on these terminological shortcuts, though. Putnam's conditions must hold as well for experiences other than pain, but may be realized by different computations in each case.

11.4. Programs

We have denoted the class of computations that can be coded with any existing programming language by $\mathcal{P}rog$. We now define Sys_0 to denote the class of systems that can run such programs. Because programs are defined relative to Instruction Set Architectures (ISA) of the underlying programming language, Sys_0 is the class of systems that can realize ISAs of existing programming languages. Denoting, as above, an ISA of a program $c \in \mathcal{P}rog$ by c_{ref}, this class is defined as

$$\begin{aligned}\mathit{Sys}_0 = \{S \in \mathit{Sys} \mid c_{\text{ref}} \subset \mathcal{C}(S) \text{ for at} \\ \text{least one } c \in \mathcal{P}rog\}\,.\end{aligned} \tag{11.4}$$

The class Sys_0 comprises all desktop and laptop computers, mobile devices, workstations, servers and supercomputers. It comprises anything that can run any Instruction Set Architecture of any existing compiler or programming language. But it does not comprise animals, or other organisms, that cannot be programmed—it does not contain organisms which are incapable of operating non-trivial logic as required by ISAs, that is. If any such animal or orgamism is conscious, the following assumption holds true.

Assumption 11.4.1. There is a system $S \notin \mathit{Sys}_0$ that is capable of conscious experience e.

In what follows, we assume computational functionalism (viz. Implication 2) and Assumption 11.4.1. The following lemma shows that if this is the case, the computation c^* is not among all programs.

Lemma 11.4.2. $c^* \notin \mathcal{P}rog$.

Proof. Assume $c^* \in \mathcal{P}rog$ and let $\tilde{S}$ denote the system in Assumption 11.4.1. Because $\tilde{S} \notin \mathit{Sys}_0$, it follows that $c^*_{\text{ref}} \not\subset \mathcal{C}(\tilde{S})$. But because $\tilde{S} \in \mathit{Sys}_e$, Implication 2 implies that $c^* \in \mathcal{C}(\tilde{S})$. This violates (11.2), so that c^* cannot be immortal. But all programs are immortal (cf. (11.3)). Hence we have arrived at a contradiction. It follows that $c^* \notin \mathcal{P}rog$. □

11.5. Turing Computation

Next, we consider Turing computations. A computation is a Turing computation iff it can be realized by (the abstract mathematical model of) a Turing machine.

Some (in fact, most) contemporary programming languages are Turing complete: they can be used to simulate universal Turing machines, meaning that they can be used to implement any Turing computation. For any Turing computation, one can write at least one program that realizes this computation, and running this program instantiates the Turing computation. This implies that

$$\mathcal{C}_{\mathrm{TM}} \subset \mathcal{P}rog\,. \tag{11.5}$$

As a consequence, we have the following lemma, which shows that the computation c^* put forward by Computational Functioanlism is not a Turing computation.

Lemma 11.5.1. $c^* \notin \mathcal{C}_{TM}$.

Proof. Follows from Lemma 11.4.2 and (11.5). □

11.6. Immortal Computation

The next lemma shows that consciousness is a mortal computation.

Lemma 11.6.1. $c^* \notin \mathcal{C}_{Imm}$.

Proof. Lemma 11.5.1 states that $c^* \notin \mathcal{C}_{\mathrm{TM}}$. Because of (11.1), we furthermore have $\mathcal{C}_{\mathrm{Imm}} \subset \mathcal{C}_{\mathrm{TM}}$. Therefore, it follows that $c^* \notin \mathcal{C}_{\mathrm{Imm}}$. □

11.7. Conclusion

We have shown that computational functionalism implies that consciousness is a mortal computation, and that consciousness cannot be a program or Turing computation. We hope that this result contributes to the understanding of computational functionalism and its implications, including questions of AI consciousness, and that it highlights mortal computation as a potential concept of interest with respect to question of the mind.

Because all contemporary Artificial Intelligence (AI) is immortal computation, the results provide initial reason to believe that no current or near-future AI can be conscious. Only artificial systems that employ mortal computations can instantiate consciousness. The results presented here do not, however, prove this to be the case. That is because it might be the case that consciousness, despite being a mortal computation, can be realized by immortal computations. Whether this is a viable option, and what precisely

it means to realize or implement a mortal computation, depends on details of the notion of mortal computation that are to be developed in future research. Because mortal computation is not Turing computation, the possibility of such realization might bear various difficulties, in case of which strong implications for synthetic phenomenology would follow.

12. Conclusion

This thesis has explored a variety of topics in consciousness science from a mathematical perspective. The goal of this concluding chapter is to review the progress that has been made, but not in the form of a synopsis of the individual projects that have been carried out—such synopsis can be found in the last section of each chapter already. Rather, this chapter attempts to paint a broader picture of the research that has happened in this PhD. It seeks to explicate both the goals and long-term visions that underlie, or have emerged from, this research.

In a sense, this chapter is an attempt to project existing results into the future. The hope is that such a projection, even if blurry, might provide the reader with a much more lively picture of the research that has begun in this PhD.

12.1. How do we Build Theories of Consciousness?

A first contribution of this PhD to consciousness science concerns the question of how theories of consciousness can, or should, be constructed.

Consciousness science, in its present stage of development, comprises a large number of theories. There are at least 39 theories published in journals that relate to the field (cf. Section 7.2.1), and likely many more in journals of different disciplines or unpublished at the present stage.

While some variation among theories might be expected given the different metaphysical assumptions and explanatory strategies that are employed (Signorelli, Szczotka, & Prentner, 2021), the bulk of the variation, arguably, may be due to the fact that there are no noteworthy constraints in proposing a theory of consciousness. Any hypothesis or experimental finding can, once singled out from its context, be presented as a new theory of consciousness. All that is required is the individuation of some property, mode, mechanism or configuration among the subject matter of the natural sciences, as well as some conjecture of how this property, mode, mechanism or configuration might relate to the conscious perception of a stimulus, the instantiation of a phenomenal property, or the subject being conscious at all, cf. Section 7.2.

This situation may be due to the fact that consciousness science does not yet have a thorough paradigm, in the sense of Kuhn (1962), for how to build a theory of consciousness. The various theories that exist are, as far as the construction of theories of consciousness is concerned, more akin to examples.[1] They embed a huge amount of

[1] I am grateful to Tim Ludwig for discussions on this topic.

very valuable insights, but a thorough paradigm that guides the construction of theories, like Newtonian mechanics in physics, for example, is not yet available.

Much of the work on theories of consciousness in this PhD was aimed at gauging the direction in which such a paradigm for the theory-building process might eventually be found. Three lines of research are related to this question, which we review in what follows.

12.1.1. The Role of Mathematics in Constructing Theories of Consciousness

A question that drove much of the research carried out in this PhD is the question of which role mathematics can play in the theory-building process in consciousness science. This includes questions like:

- **(a)** What can mathematics *add* to the theory-building process in consciousness science? Are there particular advantages when using mathematics to formulate theories of consciousness? And if so, how do these advantages pan out in practice?
- **(b)** Are there reasons for why it might be necessary to use mathematics in formulating theories of consciousness, under certain circumstances?
- **(c)** How do we actually use mathematics to build theories of consciousness?

It is important to stress that the use of mathematical formalism in constructing theories of consciousness is not obviously a good idea. The large majority of theories of consciousness that exist today is not formulated in mathematical terms, and the science is making good progress nevertheless. Any call for mathematization, either in general or in specific cases, needs to introduce good reasons for why a mathematization should be applied. Good reasons that justify not only the effort of applying the mathematization, but which also take into account the further scientific progress of theories. The exactness and devotion to detail involved in mathematical methods can easily inhibit a progressive research programme.

The following answers to the above questions are a synopsis of various research projects carried out in this PhD. On the one hand, the research projects directed at uncovering or reconstructing the mathematical structure of existing theories of consciousness, including theories that are already using formalism, like Integrated Information Theory (IIT) and Predictive Processing Theory (PP) with Active Inference, as presented in Chapters 2, 3, and 4, but also *non-formal* theories, such as Global Neuronal Workspace Theory (GNWT) and Higher Order Thought Theories (HOTT).[2] On the other hand, the research projects that work on foundations of a structural turn, as presented in Chapters 7, 8 and 9.

A more faithful account of the target phenomenon. Perhaps the most important contribution that mathematics can make to the theory-building process in consciousness

[2] The research on GNWT is in its final stages and has already been presented in the *Mathematical Spaces for Conscious Experiences* symposium at ASSC26 in NY. The research on HOTT is in its early stages. Both are unpublished at the present stage.

science is the use of mathematical structures and mathematical spaces to represent phenomenal character. This affords a more faithful representation of the target phenomenon of consciousness science in theories of consciousness, where 'faithful' is meant to indicate that the representation is true to the representandum: it resolves the details that matter.[3]

The question of how to use mathematics to represent phenomenal character and other aspects of conscious experience has taken up a large part of the research in this PhD, spanning Chapters 7 to 9. For an introduction and high-level synopsis thereof, cf. Sections 1.3 and 12.2.

What is important to note is that the introduction of mathematical spaces or structures requires one to use formal theories of consciousness: 'structural theories' of consciousness that provide a hypothesis about how the subject matter of other natural sciences (neuroscience, biology, etc.) relates to the structure of conscious experience. The application of structures of conscious experiences in theories of consciousness pulls the whole theory-building process into a new realm: the realm of mathematized theories.

Flexibility and parsimony in explicating ideas, experimental results, or metaphysics. A second advantage which mathematics might bring to the theory-building process in consciousness science is the explication of ideas, experimental findings, or metaphysics in constructing theories of consciousness.

The mathematical method is known for its dedication to detail, which does matter substantially in many circumstances. But for theories of consciousness, at least in this early stage of development of consciousness science, detail is not what matters most. It is, arguably, more important to get the broad story right than to consider all possible details.

The features of mathematics that are of advantage at the present stage of development of consciousness science are, rather, flexibility and precision. Mathematics is flexible because it allows one to define concepts precisely as intended, free of terms that have several meanings or unintended connotations. And mathematics is precise because a mathematical definition requires one to spell out all details that matter. There can be vagueness, but if so, its boundaries are precisely defined.

Because of flexibility and precision, mathematics enables one to explicate ideas, empirical findings, or metaphysics in exactly the way one would like to state them, staying true to what the idea, empirical finding or metaphysical assumption actually comprises. This is, in many cases, a large advantage over non-formal language.

Mathematics in the natural sciences. A third reason for why mathematics might, at least at some point of development, be the right language for formulating theories of consciousness is that mathematics is also the language used in most theories of natural science.

[3] For a formal definition of a faithful representation in mathematics, cf. (nLab, 2024a) and (nLab, 2024b).

Theories of consciousness are hypothesis about how the subject matter of natural sciences (usually simply called "the physical") and consciousness relate. Therefore, theories of consciousness need to refer to the subject matter of natural sciences in some form.

At the present stage of development, this reference is usually made by use of names (e.g. names of ROIs) or neuroanatomical orientations (e.g. prefrontal vs. posterior cortex). But the parts of the brain that are referred to by these names or orientations are distinguished in virtue of their structure, function, or processing. And it is at least likely that a detailed account of such structure, function, or processing requires the use of formal terms or formal models.

Movements in this direction are already taking place. Seth and Hohwy (2021)'s call for replacing the search for ROI-based Neural Correlates of Consciousness by a search "through the lens of [P]redictive [P]rocessing" (Seth & Hohwy, 2021, p. 1), and further developments in this direction in Computational Phenomenology (cf. Section 12.3), are examples of research programs that engage with mathematical models of the brain in virtue of their success in neuroscience.

The first two points mentioned above–a more faithful account of phenomenal character, and the opportunity afforded by flexibility and parsimony in explicating ideas or experimental results—constitute advantages that mathematics might bring to the theory-building process in consciousness science (question **(a)** above). The use of mathematics in the natural sciences might constitute a reason of why the use of mathematics might be necessary at some point of development of consciousness science (question **(b)** above).

There are further advantages that mathematics might afford, which deserve mentioning. They are, however, a bit more speculative, and perhaps of a more specific relevance in practice. But they constitute important ideas nevertheless.

Universality. A first advantage is tied to a desideratum for theories of consciousness which has recently been proposed by Kanai and Fujisawa (2024): that theories should be formulated in such a way that they can "determine whether a given dynamical system is conscious, irrespective of its origin or composition (e.g. whether it is a biological brain, hurricane or computer)" (Kanai & Fujisawa, 2024). This desideratum is called *universality*.

While very plausible at first look, the desideratum is not uncontested. Researchers in consciousness science whose work is more aligned with methods and theories of cognitive science argue that theories in cognitive science are not universal in this sense. These theories explain phenomena, but target only the brain. Research on the various forms of memory, say, is not required to produce theories that also hold true for AI systems. So why should a theory of consciousness?

This is certainly a fair point. For all we know, consciousness could be instantiated only in biological systems. And if it is instantiated in artificial systems, perhaps in an entirely different way.

On the other hand, one could argue that most theories of consciousness intend to say something that is true of consciousness per se, not just consciousness of systems in the scope of neuroscience. These theories seem to want to provide a universal understanding. The same holds true for views that take consciousness to be a natural kind.

Independently of whether one endorses the desideratum or not, it is arguably the case that the desideratum can only be met by theories of consciousness which are formulated in mathematical terms. This is the case because, if any language is capable of being applied to the vast range of systems that the desideratum requires, it is the language of mathematics. Any adequate description of Large Language Models (LLMs), for example, is a description in terms of mathematics: LLMs are defined in terms of formal models explicated in computer code. And hurricanes afford formal models as well. Hence any universal theory of consciousness will likely have to be a mathematical theory of consciousness.

Unification. Mathematics could, arguably, be required to unify the different theories of consciousness that exist to date.

A wonderful example of this is the model of consciousness developed by M. Blum and Blum (2021), called *Conscious Turing Machine* (CTM). This is a model of consciousness, where 'model' is understood in the sense of computer science, which is different from the notion of 'model' in neuroscience and other natural sciences. The CTM is meant to fulfil the same role in consciousness science that Turing's model of computation or Shannon's modelling of information fulfilled in computer science. A big part of the work that flows into the definition and exposition of this model rests on incorporating the core formal ideas of other theories of consciousness in that model, ranging from GNWT and IIT to theories about the evolutionary origin of consciousness, for example the proposals by Humphrey (2023).

Transcending Language. Another opportunity of mathematical approaches in building theories of consciousness that should arguably be mentioned is that they can help to overcome the confines of non-formal language. Mathematics might allow to construct theories of consciousness that rely on concepts which cannot properly be expressed in non-formal language, not because of the practical limitations of non-formal language, but because of the logical context in which non-formal language is embedded.

This idea is often important in the context of investigations that aim at formulating theories of consciousness based on Buddhist, Hinduist, or other idealist assumptions, because important concepts in these metaphysical frameworks are expressed as what looks, from a classical logic perspective, like contradictions, but nevertheless contain substantive points.

Because mathematical formalism is not necessarily tied to classical logic, or other presumptions and categories of a mind selected for in natural evolution, mathematics offers a natural and arguably unique starting point to construct theories of consciousness that embed, rely on, or are inspired by concepts that transcend non-formal lan-

guage. Mathematics, in a sense, allows one to "climb out of" the edifice erected by natural selection. Next to precision and flexibility, this may be a main reason behind the success of mathematics in foundations of physics and related disciplines.

This concludes a very initial assessment of how mathematics might be of help—or, in some cases, even necessary—in constructing theories of consciousness.

Regarding question **(c)** above—of how to go about building mathematical theories of consciousness—, the present results suggest only one important answer: there is no one-fits-all solution. Mathematical theories of consciousness are of the same type as non-formal theories of consciousness: they express hypotheses of how the subject matter of the natural sciences relates to conscious experience. The difference lies only in the language that is used to represent the subject matter of the natural sciences, conscious experiences, and the relation that holds between them. But this difference in language does not reduce options. On the contrary, it allows for more freedom in formulating theories of consciousness. Which formulation is appropriate depends on the metaphysics, idea, or empirical finding that motivates the theory. The actual work of explicating this idea will have to be done in each case separately. Making use of the methodological opportunities reviewed above in concrete cases remains challenging and requires as innovative research as always.

It goes without saying that the mathematical effort is complementary to, and not in competition with, non-formal approaches in consciousness science. Mathematical consciousness science extends and amends non-formal approaches, in the very same way as theoretical and mathematical physics extend and amend experimental physics (cf. Chapter 1). The hope is that a combination of formal and non-formal approaches can ultimately lead to a paradigm in constructing theories akin to the powerful paradigms available in other natural sciences.

12.1.2. Constraints for Theory-Building

A second line of research in this PhD that turned out to be intimately connected to the question of how to build theories of consciousness emerged from modelling experiments in consciousness science. This research led to the identification of a *constraint* for theories of consciousness, meaning: a condition that has to be satisfied. The condition is a constraint because, as we now explain, theories which violate the condition cannot actually be empirically investigated. Violation of the condition by IIT and other contemporary theories of consciousness is precisely what implies that these theories are "wrong or (...) outside the realm of science" (Doerig et al., 2019).

12.1.2.1. Constraint

The upshot of Chapters 5 and 6 is that this constraint consists in the following requirement:

A theory of consciousness must explain

why measures measure what they measure.

Here, 'measures' refers to measures of consciousness, or C-tests,[4] both of which are tools to infer information about whether a system is conscious, or what it is conscious of. 'Explain' is a shorthand for the requirement that a theory of consciousness must afford an explanation. The requirement is that a theory must explain why measures of consciousness work as intended, if the theory is true, at least for those measures that are used to test the theory's predictions.

Virtually all theories of consciousness that have been proposed to date fail to satisfy this requirement, for one of two reasons. Those theories that offer a specific and well-defined hypothesis about how conscious experience relates to the brain, such as the formal part of IIT, mostly fail because they do not model how subjective report, or other behaviour that measures and C-tests rely on, comes to reflect the content of experience as inferred by the respective measures and C-tests. For example, the specific hypotheses they propose do not contain a concrete explanation of why or how subjects can report on their experience. There is no explanation of how, when a subject reports on their experience, that report can depend on the content of experience according to the theory.

The theories that, on the other hand, operate with a less specific hypothesis about how conscious experience relates to the brain (as is the case, for example, for GNWT, which does not specify the necessary or sufficient conditions for something to count as a global workspace), fail for another reason. While the broader picture they propose often includes a sketch of a mechanism that might explain why report or other behavioural indicators come to reflect the content of consciousness, the explanations are too general to afford of an explanation of why measures measure what they measure.

As a result, in both cases, there is a theoretical possibility to vary the part of the brain that is responsible for conscious experiences, according to the theories, without changing the report. Such variations are called substitutions in (Kleiner & Hoel, 2021), presented in Chapter 5. In the case of IIT, one can, in theory, substitute a part of the system with an unfolded system, so as to change the experience while keeping the report constant. In the case of GNWT, to give another example, one can substitute the part that constitutes the workspace by a simple lookup-table system. In both cases, the experience changes, but not the behaviour. As explained in detail in Chapter 5, this leads to the problems that Doerig et al. (2019) have first spotted for IIT-like theories.

As we will see in the next section, resolving this constraint means breaking with some assumptions that are deeply embedded in consciousness science at the present stage of its development. This results in challenges for the field, most notably for theory-building, but also yields novel opportunities, in both theoretical and experimental research.

[4] The notion of C-test, proposed in (Bayne et al., 2024), was not available at the time when (Kleiner & Hoel, 2021) and (Kleiner & Hartmann, 2023) were written. But all analyses provided in Chapters 5 and 6 apply to both C-tests and measures. That is the case because C-tests have the exact same formal structure as measures; both are methodologies to infer states of consciousness from experimental data. The difference is which data is considered, and what the states of consciousness that are inferred describe.

Before discussing these implications, it should be noted that this is only a first constraint that has been discovered. It is very likely that there are more constraints on how to build theories of consciousness, which are not known at the present stage. Further research into measures and C-tests, perhaps in the form of a measurement theory for consciousness science (cf. Section 12.4), and consciousness' unique epistemic context is required.

12.1.2.2. Resolution

Resolving the constraint explained in the last section requires detailed analyses of individual theories of consciousness. Because every theory has its own proposal for how consciousness relates to the brain, it also needs its own explanation, perhaps in the form of a mechanism, of how the content of consciousness comes to determine report in just the right way. To provide an explanation that features enough details to preclude the substitutions explained in the last section is likely no easy feat; a general recipe is probably not available.

There is, however, a minimal condition that all theories have to meet, and which can serve as a starting point for investigations into how individual theories of consciousness could address the constraint. This minimal condition has first been identified in Chapter 6, in the context of the study of epistemic implications of a particular metaphysical assumption, and was refined in Chapter 10, so as to present it in a form that underlines its complete metaphysical neutrality. It is now called *dynamical relevance*.

The definition of dynamical relevance, which we will review momentarily, makes use of the fact that scientific theories of consciousness are built on top of the knowledge and insights of natural science: theories of consciousness make use of models or theories from natural sciences to express their hypotheses of what consciousness is and how it relates to the subject matter of the natural sciences. The model or theory from natural science that a theory of consciousness T is built on, or makes use of, can be called T's *reference theory*.

Dynamical relevance, then, in simple terms, is the requirement that a theory of consciousness posit that consciousness makes some difference (= is relevant) to the time evolution of the states of its reference theory. The time evolution of the states of a theory is sometimes called a theory's *dynamics*, hence the name *dynamical relevance*. Cf. Section 10.1 and Definition 10.2.1 for a more careful exposition and definition of this notion.

Dynamical relevance is a minimal condition. Any account of how a system's reports or behavioural indicators, as described by the reference theory, come to depend on consciousness as posited by the theory of consciousness, is an account of how consciousness is relevant for the system's dynamics as described by the reference theory. It is important to note that this condition is not in conflict with physicalism, but rather rests on the fact that reference theories—models of the brain, for example—express a particular state of knowledge in the sciences, which the theory of consciousness amends. Dynamical relevance could be cashed out in terms of a causal influence, but if so, between

consciousness as a physical phenomenon and the brain as a physical phenomenon. All that is required is that consciousness get a proper role in the cognitive architecture instantiated by the brain, as far as reports and other behavioural indicators of consciousness are concerned.

12.1.2.3. Implications

Because dynamical relevance is a minimal condition that must be met in order to resolve the constraint explained in Section 12.1.2.1, it can serve as a basis to analyse which implications the constraint has for consciousness science at large.

At the very least, the constraint requires a fundamental change in theoretical thinking in the context of theories of consciousness. In addition to the "from brain to consciousness" direction, it requires us to engage in the "from consciousness to brain" direction as well, so as to specify which difference consciousness makes in our understanding of brain function.

But it would be wrong to think that these implications merely affect theories. Rather, what the constraint, and the analysis in Chapter 5 show, is that we need to step away from conceiving theories and measures as independent from each other. Rather, we must develop a coherent perspective on testing theories of consciousness, in which both the theory's prediction and the measure's function are considered together. Both are part of understanding, and deriving, predictions in consciousness science, and they cannot be separated. We need a comprehensive perspective of testing theories, which manages the "delicate navigation (...) between Scylla and Charybdis" (Kleiner & Hoel, 2021) in this context.

12.1.3. Structural Theories of Consciousness

A third contribution of this PhD thesis to the question of how to build theories of consciousness concerns the use of mathematical structure and mathematical spaces in theories of consciousness.

One important advantage of using structure and spaces for building theories of consciousness has already been mentioned in Section 12.1.1: structures and spaces enable theories of consciousness to represent phenomenal character more faithfully. This might constitute a desideratum for theories of consciousness in its own right, but the use of structure also has practical—and quantifiable—consequences that are of independent interest: a larger explanatory scope and an increase, everything else being equal, in predictive power.

Structural theories of consciousness have a larger explanatory scope than their non-structural counterparts, because structures and spaces represent phenomenal character more comprehensively than non-structural approaches can do. They represent the various qualities or phenomenal properties that are instantiated in single experiences, but they also represent phenomenal relations that hold between them. Furthermore, structures and spaces can represent more features and more details of phenomenal

character than would otherwise be possible. As a consequence, structural theories can explain more of phenomenal character than their non-structural counterparts are capable of.

Structural theories of consciousness have, everything else being equal, a larger predictive power than their non-structural counterparts, because a structural theory of consciousness constitutes a much more detailed and rigorous hypothesis of how conscious experience relates to the brain.

Consider, for example, Global Neuronal Workspace Theory (GNWT). GNWT does not, in its current formulation, offer an account of how phenomenal character is determined, it only explains how signals from parallel processors enter consciousness. But a promising idea about the relation between GNWT and phenomenal character—that it is the content of the workspace that determines phenomenal character in full—is available in the field. Structural tools allow to turn this idea into a scientific hypothesis, which results in a range of additional predictions that could, in principle, be tested.

The exploration of structural tools for building theories of consciousness has just begun. A big part of this PhD, presented in Chapters 7 to 9, was devoted to providing a solid foundation for the use of structure and spaces in scientific theories of consciousness.

12.2. Promises and Foundations of a Structural Turn

Over the previous years, consciousness science has seen a steep increase in the use of mathematical structures and mathematical spaces to describe or represent consciousness. These developments might constitute early signs of a structural turn in consciousness science, in which mathematical spaces and structures are used, in conjunction with the tools that are available already, to improve theories, experiments, measures and concepts.

Much of the work carried out in this PhD was directed at understanding the foundation of structural research, and the opportunities a structural turn might afford. The upshot is that structural approaches have a huge potential in consciousness science. If implemented in the right way, they will lead to better theories, better experiments, and better concepts. They carry a promise to vastly extend the range and scope of consciousness science, and might offer a new perspective on many questions currently studied in the field.

In what follows, we explain some of the more palpable promises of a structural turn. Needless to say, these promises can only be realised by community efforts, spanning researchers across different fields and research programmes.

12.2.1. Structural Theories

A first big impact that structural approaches can have on consciousness science concerns theories of consciousness, specifically the use of mathematical structures or spaces to represent conscious experience and phenomenal character in theories of consciousness, as mentioned already in Section 12.1.1.

Such 'structural theories' are not different in kind from binary theories of consciousness. Structural approaches do not constrain the metaphysical or conceptual content of theories. They too are hypotheses about how consciousness and the subject matter of the natural sciences relate. The difference between structural and non-structural theories is that the former employ a different way to handle, describe or represent conscious experience, based on mathematical spaces and mathematical structure.

We have already touched on the advantages that structural theories might bring to consciousness science in Section 12.1.3: structural theories are *more explanatory*; and structural theories are *more predictive*, cf. Chapter 7.

Further work is necessary to realise these advantages. At the present stage of development, only three structural theories are available: Integrated Information Theory (Oizumi et al., 2014), some Higher Order Thought Theories (Brown, in press) and Expected Float Entropy Theory (Mason, 2013). But while all of these theories employ or account for *some* structure to represent phenomenal character, they do not yet account for the *actual* structure as found in psychophysical approaches and mathematical phenomenology. Only when they do this will some of the advantages start to realise.

What this PhD contributes to this promising strand of developments is work on the foundations of what it means to represent or describe phenomenal character in terms of mathematical structure in Chapter 8 and 9, as well as a critical assessment of some of the ideas that have started to emerge on how to build structural theories of consciousness in Chapter 7.

12.2.2. Structural Experiments

A second huge promise of a structural turn in consciousness science concerns experiments. Making use of structure to describe or represent conscious experience is likely to change measurement in consciousness science.

A case in point is the measurement of Neural Correlates of Consciousness (NCCs). As shown by (Fink et al., 2021), if certain assumptions about structure hold true, most notably (a) structuralism—the idea that the structure of conscious experiences fully determine all non-structural experiential facts, including which phenomenal properties are instantiated in an experience—, and (b) that there is an isomorphism between the physical structure of the brain and the phenomenal structure of experience, a whole new paradigm to search for the NCCs can be provided. This paradigm might not rely on reports in near-threshold contrast conditions (Baars, 1986).

Unfortunately, as shown in Chapter 7, the assumptions presumed in (Fink et al., 2021) are not justified as general conditions on which an NCC research programme can rely. Most notably, the assumption of an isomorphism between the physical and phenomenal domains, or of a structure-preserving mapping more generally, does not serve the purpose it is required to serve in this context. Therefore, the research programme outlined in (Fink et al., 2021) can only be understood as a research programme that presumes a specific class of structural theories of consciousness; this class is to all structural theories of consciousness what the class of identity theories is to all non-structural theories

of consciousness.

In spite of these constraints, the general point that Fink et al. (2021) are in essence making—that structural approaches might afford a whole new class of measurement schemes for NCCs, and perhaps also for other measurements in consciousness science, one might add (cf. Section 12.4), holds true. The identification of technical difficulties of this first proposals is a sign of good progress, and might lead to ways of overcoming them. The future of NCC research, and measurement of consciousness in general, might well lie in methodologies that combine structural tools with novel experimental or philosophical ideas.

12.2.3. Structural Concepts

The most exciting promise of a structural turn in consciousness science, which (of course) is also the one which is most difficult to assess at the present time, might arguably be the creation of new methodologies for consciousness science, where the term 'methodology' is used in the general sense of a "body of methods used in a particular field of study or activity" (Oxford English Dictionary, 1989). This includes theories of consciousness and novel experimental tools, as discussed in the previous two sections, but may also go beyond them.

Structural approaches may offer entirely new avenues for conceptual engineering. This could be the case, for example, in the context of mathematical phenomenology (cf. Section 12.3), where structural approaches could afford entirely new ways of representing, describing and thinking about aspects of phenomenology. Perhaps mathematical structural approaches can address those aspects of phenomenal experience that are difficult to express in common language, for example nondual awareness or nonegoic reflexivity (Metzinger, 2024), to name just two. And it could be the case in psychophysics, where structural approaches could afford entirely new ways of representing and measuring structural phenomenal properties, new ways which are grounded in the mathematical structure of said properties.

These opportunities are particularly interesting from an illusionist or discourse eliminativist perspective (Frankish & Sklutová, 2022; Irvine & Sprevak, 2020), both of which hold that existing concepts that address the target phenomenon of consciousness are misleading, and should either be discarded or regarded as illusory. However, they do not aim to discard the field of consciousness science entirely, but rather propose alternative concepts. "The positive part of a discourse eliminativist's argument aims to show that an alternative way of talking, thinking, and acting is available" (Irvine & Sprevak, 2020).

Structural concepts might offer such an alternative way of talking, thinking, and acting. This is the case because structural concepts can be grounded directly in empirical data, much like is the case in the monumental foundational measurement theory of Krantz et al. (1971). Because of this, structural methodologies could provide a foundation to develop concepts and methods that overcome what appears to be—from the perspective of these views—ill-founded conceptual foundations.

The methodology for structural representations of conscious experience developed

in Chapters 8 and 9 of this PhD thesis offers a starting point for exploring these opportunities in more detail.

12.3. Mathematical Phenomenology & Computational Phenomenology

The term 'phenomenology' denotes various concepts in consciousness science. It is sometimes used to denote the object of investigation in consciousness science, viz. what is also denoted by terms like 'conscious experiences' or referred to by locutions like "what it is like" (Farrell, 1950; Nagel, 1974). But it also denotes a way of engaging with consciousness scientifically and philosophically. Phenomenology in this latter sense refers to a discipline (and movement) in philosophy that contributes to consciousness science, but also has goals that transcend it. It is usually taken to be grounded in the works of Edmund Husserl, Martin Heidegger, Jean-Paul Satre, and Maurice Merleau-Ponty, among others.

Phenomenology has important insights, and important methodologies to offer to consciousness science. In contrast to conceptions of consciousness that are prominent in other disciplines and other parts of philosophy, phenomenology emphasises the lived character of experience, how experience constitutes itself, and structural aspects of a more dynamical nature. All of these are part of the object of investigation of consciousness science, and a full scientific understanding of consciousness will have to include these aspects as well.

12.3.1. Mathematical Phenomenology

Mathematical phenomenology, also called mathematized phenomenology, aims to apply mathematical concepts and techniques in phenomenological investigations, most notably in the form of mathematical presentations of the results of such investigations. Mathematical phenomenology has been pioneered by Petitot (1999) and Yoshimi (2007). Important recent work has been carried out by Prentner (2019, 2024b), who uses mathematical tools related to pre-topologies to provide a mereological account of the unity of consciousness, intentionality, the self-world distinction, and time.

The research carried out in this PhD does not contain a project devoted to mathematical phenomenology, but mathematical phenomenology is both a future outlook and important background for research carried out here: the research on mathematical structures of conscious experience, presented in Chapter 8 is aimed at mathematical phenomenology as much as it is aimed at psychophysical spaces. It offers a definition and methodology for how to use mathematics to represent or describe conscious experience, which is based on variations. These variations can result from the variations of stimuli that are presented to a subject, as in the case of psychophysical spaces, but they can also result from imagined variations in the very sense of Husserl's eidetic variations (Husserl, 1939; De Santis, 2011).

Because the definition and methodology can be applied to both variations of stimuli and eidetic variations, the research presented in Chapter 8 offers the possibility of connecting psychophysics with phenomenology. It offers the hope of unifying quality spaces as constructed "in the lab" in psychophysical measurements with mathematical representations of phenomenology as constructed in phenomenological studies. This might lead to a more comprehensive and thorough representation of conscious experiences in terms of mathematical structure, and offers a hope of cross-inspiration of the two fields.

Whether or not these hopes realise in practise is a matter of future study. Only preparatory work has been undertaken in this PhD. But it is inspiring to think that mathematics might be the key to arrive at a more unified and integrated agenda in consciousness science, which transcends the mutual criticisms of methodologies that exist at the moment.

12.3.2. Computational Phenomenology

A particularly interesting development in the context of mathematical approaches to the mind is computational phenomenology. Computational phenomenology, in its original conception, is the modelling of phenomenology in terms of computational tools from computer science (Harlan, 1984). More recently, the term has been used to denote the modelling of phenomenology in terms of the computational framework provided by Predictive Processing Theory and its Active Inference doctrine (Ramstead et al., 2022).

Because models of computation are mathematical models, computational phenomenology can be seen as a part of mathematical phenomenology. And much like mathematical phenomenology in general, computational phenomenology in particular requires a solid foundation of how to represent phenomenology in mathematical terms. Foundational work on what it means to represent phenomenology mathematically is also foundational work of what it means to represent phenomenology computationally.

In light of this connection between mathematical phenomenology and computational phenomenology, computational phenomenology presents an interesting future application of some of the work on mathematical structures of conscious experience presented in Chapters 7 to 9.

Another noteworthy relation between the work carried out in this PhD and computational phenomenology is research on the mathematical structure of Predictive Processing Theory (PP) and Active Inference, presented in Chapter 4. This chapter provides an account of the mathematical structure of PP and Active Inference models in the mathematical language called category theory. This is not simply a reformulation of the formulae in which the theory is usually defined, but an account of which mathematical structure lies behind these formulae.

The hope behind the research in Chapter 4 was that the mathematical account of PP and Active Inference can be helpful in understanding how the theory relates to conscious experience. This hope carries over to recent developments in computational phenomenology.

12.4. What is Measurement in Consciousness Science?

Contrary to public conception, consciousness can be measured. There are several so-called *measures of consciousness* that can be used to measure whether a stimulus has been perceived consciously in experimental trials, cf. e.g. (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). And there are empirical tests that can be applied to humans and some non-human animals to investigate whether they are capable of having conscious experiences, or not, now called *C-tests* (Bayne et al., 2024). Cf. Section 1.2 for an introduction to both.

What there isn't, at present, however, is a substantive *theory of measurement* (also called *measurement theory*) for consciousness science that provides a foundation for measurements of the various forms.

Measurement theories are an important part of those sciences where measurement is not straightforward. Consider, as an example, the case of psychology. In the first part of the 20th century, the question of whether there is measurement at all in psychology has been heavily debated, so much so that in 1932 a committee of the British Association for the Advancement of Science was appointed "to decide whether or not there was such a thing as measurement in psychology" (Borsboom, 2005), cf. (Ferguson et al., 1940). The committee's report was highly divided, with a majority of members around the physicist Norman Campbell strongly rejecting claims about the possibility of measurement in psychology.

In response to this rejection by part of the committee, psychologists started to develop theories of measurement that are targeted specifically at psychological experiments, first in the form of scales by Stevens (1946) and then in the form of axiomatic theories of measurement (Cliff, 1992), the most well-known of which is foundational measurement theory (Krantz et al., 1971), also called representational measurement theory. These developments were pivotal to the progress of psychology in the 20th century (Michell, 1999; Borsboom, 2005), and still form the basis of much of the empirical work that is being carried out.

Another example of where a theory of measurement was crucial for understanding the intricacies of measurement is quantum theory in physics. Quantum theory comprises a comprehensive account of measurement, where complicated measurement apparati, that might fill a whole room in a lab, are represented by comparably simple mathematical objects: self-adjoint operators on Hilbert spaces,[5] which in finite-dimensional cases, and given a choice of basis, can be thought of as matrices over complex numbers. Based on these mathematical representations of measurement devices, quantum theory provides an account of how measurement interfaces with the time evolution of a system, which includes an account of possible results of the measurement procedure, as well as an account of how the measurement procedure changes or modifies the state of the system. While this account of measurement is also the source of the notorious measurement problem of quantum theory (cf. e.g. (Myrvold, 2022)), it is hard to imagine which pro-

[5] In present-day quantum information, measurements are represented by the more general concept of *quantum instruments*, developed by Ozawa (1984).

gress could have been possible without the introduction of this part of the theory by von Neumann (1932).

Given consciousness' unique epistemic context, it is likely that a theory of measurement will be as consequential for consciousness science as it has been for psychology and quantum physics. Building on these examples, a preliminary list of desiderata for a measurement theory of consciousness science could comprise the following elements:

(a) An abstract representation of measurement. Both psychology and physics, despite being different in many respects, rest on abstract (and, in fact, mathematical) representations of measurement procedures. Correspondingly, a measurement theory for consciousness science will likely have to comprise abstract representations of measurements in consciousness science. Such representations should be both descriptive (in the sense that they are built on and represent the actual empirical measurements that are carried out) and normative (in the sense that they guide the development of novel forms of measurement).

(b) What constitutes a measurement, and what not? The abstract representations of measurement should account for what measurement of consciousness is, and delineate between what counts as measurement, and what does not. For example, does simple verbal report constitute a measurement? And how far can no-report paradigms be pushed while still constituting measures of consciousness?

(c) Which aspects of consciousness are measurable, and which not? A theory of measurement for consciousness science should provide conditions that have to be in place for some aspect (or property, mode, part, etc.) of conscious experience to be subjected to measurement.

(d) How does measurement integrate with theories of consciousness? A theory of measurement should furthermore account for how measurement integrates with theories or models of consciousness. In fact, lack of such integration is precisely what plagues the theory-building process in consciousness at the present stage of development, cf. Section 12.1.2.

(e) Which conditions make measurement possible? The abstract representations of measurement should, ideally, also provide conditions that make measurement possible. This is the case, for example, for representational measurement theory in psychology, which relates measurement of a qualitative structure to representation and uniqueness theorems regarding quantitative representations of that structure.

Because measurement theory in both psychology and physics is based on a mathematical representation of the measurement process, it is at least prima facie possible that mathematics might offer the right tools to construct a measurement theory for consciousness science as well. Whether or not this is possible, or fruitful, remains an open question, and scepticism is fully warranted. But in light of the importance (and, in some cases, necessity) of mathematical theories of measurement in other sciences, the possibility seems worth exploring.

Within this PhD, some progress has been made in regard to such exploration of points **(a)** and **(d)**, albeit in unsystematic form. Regarding **(a)**, a first mathematical representation of measures of consciousness, which also applies to C-Tests, has been developed in Chapter 5. This representation has subsequently been applied to investigate problems and necessary conditions of the integration of measures of consciousness in theories of consciousness as required by **(d)**, cf. Section 12.1.2 for a summary. Furthermore, the work on foundations of structural methodologies in Chapters 7, 8, and 9, might contribute to exploring **(c)** when combined with ideas from axiomatic measurement theory.

It is needless to say that these results constitute only very initial steps, and a further exploration of all of the above-mentioned questions is urgently needed. The hallmark of a successful measurement theory, the allocation of new measurement procedures to a field, as for example the case with additive conjoint measurement (Luce & Tukey, 1964) in psychology, is still nowhere in sight.

12.5. No-Go Theorems in Consciousness Science

An important methodological tool in physics are so-called *no-go theorems*. These are theorems, in the mathematical sense of the term, that establish a conclusion about a subject matter of interest based on mathematical assumptions and a proof. The conclusion usually establishes that something is impossible, hence the 'no-go' in the name. A well-known example in physics is John Bell's proof that local realism—roughly, in this context, the idea of a world composed of localised elements with definite states—cannot be true if quantum theory is true (Bell, 1964).

The idea that no-go theorems might be useful in consciousness science goes back to Ryota Kanai. Inspired by this idea, some of the research in this PhD made use of the methodology of no-go theorems, explicitly in Chapters 10 and 11, but implicitly also in Chapter 5.

Fundamentally, the idea behind research based on no-go theorems in consciousness science is the very same as in physics: to make use of assumptions that are mostly uncontested, so as to derive, with mathematical rigour, a proof of a statement which "is a bombshell—hardly anyone would have guessed" (Edgington in response to (D. Lewis, 1976), quoted in (Leitgeb, 2013)), ideally speaking.

In the case of Chapter 10, for example, the assumptions concern general details related to the design and manufacturing of the chips on which contemporary AI systems run (CPUs, GPUs, etc.), as well as the assumption that consciousness is dynamically relevant. The somewhat unintuitive result that the no-go theorem establishes is that contemporary AI systems cannot be conscious. In Chapter 5, to give another example, the assumptions concern the general mathematical form of contemporary neuroscientific theories of consciousness, as well as the relation between theories of consciousness and measures of consciousness. The somewhat unintuitive result that follows is that theories of consciousness can always be falsified, cf. Section 12.1.2.

Formal theorems are of course well-known in philosophy under the banner of math-

ematical philosophy, cf. Section 12.7, and the application of formal methods and formal theorems carries as much potential for consciousness science as it does for philosophy. Most notably,

> "it forces us to put our cards on the table, that is, to make tacit presuppositions explicit; it helps us to separate the essential from the accidental by making transparent what exactly is needed to make an argument go through; where two areas of philosophy share enough mathematical structure, it may allow us to translate arguments in the one area into arguments in the other; it functions as a means by which we can put some of our "intuitions" to the test and correct our epistemic biases (...); it facilitates the illustration of abstract circumstances by means of diagram (...), it forges unexpected connections from philosophy to those scientific areas in which mathematical methods are accepted as a standard anyway" (Leitgeb, 2013, p. 274); and it allows for the whole suite of automated deduction to be applied to a field.[6]

In the case of no-go theorems, in addition to establishing a particular conclusion, no-go theorems also serve a second purpose in scientific methodology, a purpose which is somewhat implicit in physics but which should be made explicit in consciousness science: they shift attention and resources from the subject matter addressed in the conclusion of the theorem to the subject matter addressed by the assumptions. The hope in using no-go theorems in consciousness science, therefore, is also to shift attention, and potentially resources, to the assumptions that feed into a theorem.

A major conclusion of Chapter 10, for example, is that more attention needs to be placed on studying the substrate on which AI systems run. A major conclusion of Chapter 11 is that more research is needed to understand the novel concept of computation that is emerging at the present time; and the conclusion of Chapter 5 is that the present paradigm of formulating and testing theories of consciousness needs revision, cf. Sections 12.1.2.

Mathematical methods and formal theorems have been very useful in philosophy, and no-go theorems have made a large impact to the development in physics. It is likely that they can also play a noteworthy role in making progress in consciousness science, and it would be nice to see further explorations of this opportunity.

12.6. Artificial Consciousness

Investigations of the potential of Artificial Intelligence (AI) to exhibit conscious experiences, and of the nature of those experiences where they are indeed possible, are starting to become a key area of research in consciousness science. The list of questions that this area of research will have to answer is long. It comprises, for example, the

[6] I would like to thank Stephan Hartmann for pointing me to the latter.

following questions:[7]

1. **Are AI systems conscious?** – Or to be more precise, which AI systems are conscious? This is the question of whether AI systems can have conscious experiences at all, independently of what the experiences are.
2. **Are there *tests* for whether AI systems are conscious?** – This is the question of whether there are operational procedures that can be applied to AI systems so as to infer whether they are capable of having conscious experiences. As of late, such procedures have come to be called *C-tests* (Bayne et al., 2024). Simple examples, like direct interpretation of what Large Language Models (LLMs) state about their own experiences, are not suitable for rigorous tests, simply because LLMs are trained on huge amounts of data that include analyses of and statements about consciousness, so that a suitable prompt, given the appropriate fine-tuning, can lead to corresponding reports independently of whether LLMs are conscious or not. Other tests will have to be found.
3. **Are there theoretical means to assess whether AI systems are conscious?** – Here, 'theoretical' includes both scientific and philosophical methods. A particularly important question in this class is:
4. **Can scientific theories of consciousness, or models of consciousness, provide reliable assessments of consciousness in AI systems?** – The emphasis here is on 'reliable'. Theories of consciousness are hypotheses about how conscious experience and the subject matter of the sciences relate. The question is whether a theory of consciousness that targets the subject matter of the brain sciences can also be applied rigorously to AI systems. Are the theoretical constructs, and the empirical evidence, rigorous and detailed enough to warrant the application of theories of consciousness to AI systems?
5. **Which conscious experiences do AI systems have, when conscious?** – This is the question of the phenomenal character of the conscious experiences of AI systems—the question of what it is like to be an AI system in a particular state. Are the conscious experiences of AI systems anything like human conscious experiences? If so, what are the differences? If not, is there anything that can be said about AI's experiences? This includes, in particular, the following question:
6. **Can AI systems feel pain?** – Or do they otherwise suffer? This is an important ethical question, without resolution of which there is a potential for humanity to create a tremendous amount of suffering, cf. (Metzinger, 2021).
7. **Artificial Phenomenology** – Is it possible to apply the basics of Phenomenology to artificial systems, so as to develop an understanding of the phenomenology of artificial systems, if such phenomenology exists? Perhaps by use of mathematical phenomenology (Section 12.3), or objective phenomenology (Nagel, 1974; Lee, 2021)?

[7] I would like to cordially thank Lenore Blum and Ryota Kanai for discussions on the topic. The list of questions presented here came up in a discussion with them when preparing for the ASSC27 *Blueprints for Machine Consciousness* symposium in Tokyo.

8. **Are there measures of artificial consciousness?** – This is the question of whether it is possible to construct measures of consciousness that can be applied to AI systems, for example to find out whether AI systems experience a particular stimulus consciously, or not.
9. **Can one build conscious AI systems?** – Do the theories, or tests, provide enough details to create blueprints for building conscious AI systems? Can we implement the precise properties that theories of consciousness pin down as sufficient for consciousness? And if so, do they provide enough details to warrant assessments of the type of conscious experience the AI system will have? Can it be ascertained, for example, that the AI system will not be in pain or constantly suffer, as required by (Metzinger, 2021)?
10. **Which forms of computation are suitable to support artificial consciousness, if any?** – This question comprises two questions: the question of which types of computation can support consciousness—can a Turing computation, say? And the question of which specific properties a computation would have to exhibit so as to support artificial consciousness, if any does.
11. **Which role did evolution play in the emergence of conscious systems?** – And which implications does this have for consciousness of artificial systems?
12. **Does consciousness matter for existential risk?** – This is the question of whether conscious AI systems, in particular self-conscious AI systems, pose a particular worry in the context of alignment and existential threats.

Research within this PhD has contributed to questions **3.**, **10.**, and **11.**. Concerning question **3.**, the PhD offers a perspective on the question of AI consciousness that focuses on the actual substrate of contemporary AI systems: CPUs, GPUs, and other processing units. The research, presented in (Kleiner & Ludwig, in press) and Chapter 10, lead to a no-go theorem that speaks against the possibility of artificial consciousness for systems that run on contemporary chips. Cf. Section 1.4.1 for an introduction to and summary of this research.

A contribution to questions **10.** and **11.** can be found in Chapter 11, which is concerned with the distinction between mortal and immortal computation, introduced by Hinton (2022). This distinction is related to (but does not precisely track) the distinction between the types of computation carried out by systems that have naturally evolved and the types of computation carried out by contemporary AI systems. Chapter 11 finds that, perhaps surprisingly, computational functionalism in its original conception by Putnam (1967), sides with mortal computation, rather than immortal computation. This constitutes another argument against the possibility of consciousness in contemporary AI systems. The emphasis, however, lies on 'contemporary', as developments that might transcend the paradigm of computation that has shaped the last half century are already underway. Cf. Section 1.4.2 for a summary of the research presented in Chapter 11.

12.7. Mathematical Philosophy of Mind

Mathematical philosophy is "the application of mathematical methods to philosophical questions and problems" (Leitgeb, 2013, p. 269). *Mathematical philosophy of mind*, correspondingly, is the application of mathematical methods to philosophical questions and problems of the mind.

Following Leitgeb (2013)'s explication of the role, tasks and opportunities of scientific philosophy, of which mathematical philosophy is a part, mathematical philosophy of mind can be understood in three equally valid ways. Firstly, it can be understood as *philosophy for the mind sciences*. That is, a philosophy that reflects on the developments in the mind sciences which are mathematical in nature, on a meta-level, with the goal of reforming or improving those developments. Secondly, it can be understood as *philosophy that is part of the mind sciences*. That is, a philosophy which works hand-in-hand with scientists were mathematical methods appear, are applied, or could be helpful in the mind sciences. According to this understanding, mathematical philosophy of mind works with the same object languages as the sciences of the mind, but addresses some of the more general and more fundamental mathematics-related questions that appear. Thirdly, mathematical philosophy of mind can be understood as *philosophy of mind done with mathematical methods*. According to this understanding, mathematical philosophy of mind targets the questions, problems and hypotheses that are unique to philosophy of mind, but uses formal and mathematical methods to do so, for example formal explications of non-formal concepts. Mathematical philosophy of mind, according to this last understanding, does not aim to improve progress in the mind sciences, but rather is primarily concerned with the long-standing questions of philosophy of mind.

The research carried out in this PhD clearly falls into the second category, it is mathematical philosophy of mind carried out as part of the mind sciences. Inspired by Metzinger (2007)'s analysis of the types of interaction between philosophy of mind and the mind sciences, this research can be classified as contributing to the following four pillars of mathematical philosophy of mind.

Analysis of the target phenomenon. An important role of philosophy of mind in relation to the sciences of the mind, and correspondingly of mathematical philosophy of mind, can be subsumed under the header of analysis of the target phenomenon. This includes, mainly, the study of concepts that refer to the target phenomenon, so as to make the phenomenon, or properties thereof, accessible to scientific investigations. Such analysis can include both work on existing concepts and the proposal of new concepts, and aims to improve progress in the sciences either by making these concepts available for use in scientific investigations, or by helping scientists to conceptualise the problem under consideration.

Pivotal examples of such analyses in consciousness science are the introduction and analysis of qualia (C. I. Lewis, 1929; Peirce, 1866; Dennett, 1988; Shoemaker, 1991; Block, 2004), of phenomenal consciousness (Chalmers, 1996; Husserl, 1960), of the 'what it is like to be' characterisation of experience (Farrell, 1950; Nagel, 1974), and of

the hard problem of consciousness (Chalmers, 1995) and the notion of an explanatory gap (Levine, 1983). The last two have played a particularly important role in shaping what scientists think about their object of investigation, independently of whether they are in agreement with these analyses or not.

Mathematical philosophy of mind is particularly suitable to carry out such analyses in cases where formal concepts are applied to the target phenomenon. In the context of consciousness, this is the case for mathematical representations of conscious experience or phenomenal character. Research in this PhD has analysed existing proposals of such representations in Chapters 7 and 8. It has identified a number of shortcomings, and proposed a new definition of such representation in Chapters 8 and 9. The hope is that this new definition is helpful to make progress in structural research in both theory and experiments in consciousness science.

Analysis of methods. A second pillar of how philosophy of mind supports the mind sciences is the analysis of methods used by the latter, including critical analysis of existing methods, so as to investigate whether these methods work as intended, constructive analyses of existing methods, so as to improve these methods, and the proposal of new methods. Important examples in consciousness science are the critical analysis of measures of consciousness in (Irvine, 2012) or (Michel, 2019) and the constructive analysis of C-Tests in light of natural kinds in (Bayne et al., 2024).

Mathematical philosophy of mind offers two avenues for extending this work. On the one hand, it can cope with methods in the sciences which are thoroughly mathematical in nature. On the other hand, it can apply mathematics to provide new analyses of methods of the sciences which are not mathematical in nature.

This PhD contains an example of each kind of contribution. Regarding analyses of mathematical methods in consciousness science, it offers an analysis of a new method to search for NCCs which has been proposed based on structuralist assumptions, cf. Chapter 7 and Section 12.2.2. The upshot is that this new method uses assumptions which are not justified in this context. Regarding applications of mathematics to analyse methods which are not mathematical in nature, it offers an analysis of the contemporary paradigm for testing theories of consciousness in Chapters 5 and 6, cf. Section 1.3 for an introduction and review of these results.

Analysis of results. A third pillar of how philosophy of mind interacts with the mind sciences is the analysis of results of investigations of the mind sciences. This includes results of empirical studies as well as results of theoretical investigations, both of which comprise proposed theories of consciousness. The goal is to improve the understanding of such results, for example by critical analysis of whether the conclusions that are being drawn are justified, or by constructive analyses of what the result might mean or imply. Examples of this mode of interaction in the case of consciousness science are the numerous analyses and criticisms of theories of consciousness, and the analyses of implications of experiments for the larger questions of the field.

Mathematical philosophy of mind can extend this interaction to results which are thoroughly mathematical in nature. The main examples thereof, to date, are formal theories. Analysis and improvements of two main formal theories in the field are presented in Chapters 2 and 3 (Integrated Information Theory) and 4 (Predictive Processing and Active Inference).

Bottom-up constraints. A fourth mode of interaction between philosophy of mind and the sciences of the mind, which Metzinger (2007) identifies, concerns the use of scientific results as bottom-up constraints on philosophy of mind, so as to inform philosophical research on philosophical questions, for example metaphysical theories. Prime examples in consciousness science are the many philosophical studies of neurological disorders of consciousness, such as blindsight (Cowey & Stoerig, 1991; Stoerig, 2006), agnosia (Devinsky, Farah, & Barr, 2008), dissociative identity disorders (Kihlstrom, 2005) or neglect (Bisiach, Luzzatti, & Perani, 1979).

This mode of interaction is more aligned with the third way of understanding mathematical philosophy mentioned above, and there are no thorough examples of such interaction in this PhD. Perhaps the formal exposition of structuralist assumptions in Chapter 7 (cf. Section 7.3), and the assessment of such assumptions in light of the types of structures that appear in the sciences could be taken to be an example, but knowledge about the mathematical structures of conscious experience is too limited at the present point of development to draw any thorough metaphysical conclusions.

The above classification of mathematical philosophy of mind leaves out Chapters 10 and 11. Perhaps that is the case because they are more aligned with the *philosophy of mind done with mathematical methods* understanding of mathematical philosophy. But the question of artificial consciousness, which these chapters target, is also an important question of consciousness science. Perhaps it is best to consider philosophy of mind as part of consciousness science, and avoid making a strict distinction altogether.

Mathematical philosophy has already achieved a huge progress in epistemology, philosophy of mathematics, metaphysics, ethics, social philosophy and philosophy of language (Hartmann & Sprenger, 2012; Leitgeb, 2013). As consciousness science is reaching the stage where mathematical tools and methods become more widely applicable, mathematical philosophy of mind might too. The big promise of mathematical philosophy of mind to consciousness science is that it can help facilitate the important interaction between philosophy and consciousness science in the novel phase of mathematized consciousness research.

The exploration of mathematical philosophy of mind, therefore, is likely a worth-while enterprise. “The desiderata of exactness and fruitfulness will always ‘pull’ explication towards the application of mathematical methods” (Leitgeb, 2013, p. 272).

12.8. Synopsis

The research carried out in this PhD contributes to experimental, theoretical, conceptual, and methodological questions in consciousness science (Chapter 1). It showcases that mathematical methods and mathematical tools can be helpful to consciousness science in tasks as diverse as constructing theories of consciousness (Section 12.1), realising a structural turn (Section 12.2), exploring mathematical and computational phenomenology (Section 12.3), understanding measurement in consciousness science (Section 12.4), applying no-go theorems (Section 12.5), studying the question of artificial consciousness (Section 12.6), and realising the promises of a mathematical philosophy of mind (Section 12.7).

The goal of the present chapter was to project the results of this PhD into the future, so as to provide the reader with a broader picture of what the research in this PhD means, how it could be further developed, and how it might factor into the future development of consciousness science.

The research carried out in this PhD thesis is grounded in the conviction that *every single theory, experiment, and concept* in consciousness science *is proposed for some good reason and embodies some good idea*. Consciousness science is still in the early stages of its development, and as the field continues its journey, all of these ideas and reasons can make a decisive difference to further progress. Perhaps the novel developments in the field, a small part of which has been surveyed in this chapter, will rely on precisely some such idea or reason.

It was a huge pleasure and honour to be part of the late stage of the early development of consciousness science during this PhD. "We are still at the beginning–the best is yet to come."[8]

[8] This is a quote from (Leitgeb, 2013), though out of context. Leitgeb (2013) applied it to mathematical philosophy; here it is meant to apply to consciousness science.

Bibliography

Aaronson, S. (2014). Why I am not an Integrated Information Theorist. *Shtetl-Optimized: The Blog of Scott Aaronson*.

Abramsky, S., & Coecke, B. (2004). A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.* (pp. 415–425).

Alais, D., Cass, J., O'Shea, R. P., & Blake, R. (2010). Visual sensitivity underlying changes in visual consciousness. *Current biology*, *20*(15), 1362–1367.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., … others (2023). Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology*, *19*(10), e1011465.

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Computational Biology*, *10*(12), e1003966.

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2017). What caused what? An irreducible account of actual causation. *arXiv:1708.06716*.

Albantakis, L., & Tononi, G. (2019). Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy*, *21*(10), 989.

Alkire, M. T., Hudetz, A. G., & Tononi, G. (2008). Consciousness and anesthesia. *Science*, *322*(5903), 876–880.

Association for Mathematical Consciousness Science. (2023). *The Responsible Development of AI Agenda Needs to Include Consciousness Research.* Open Letter.

Atmanspacher, H. (2020). The Pauli–Jung conjecture and its relatives: A formally augmented outline. *Open Philosophy*, *3*(1), 527–549.

Baars, B. J. (1986). What is a theory of consciousness a theory of?–The search for criterial constraints on theory. *Imagination, Cognition and Personality*, *6*(1), 3–23.

Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*(4), 292–309.

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

Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology*, *4*(6), e1000091.

Ball., P. (2019). Neuroscience Readies for a Showdown Over Consciousness Ideas. *Quanta Magazine*.

Barrett, A. B. (2014). An integration of Integrated Information Theory with fundamental physics. *Frontiers in Psychology*, *5*, 63.

Barrett, A. B., & Mediano, P. A. (2019). The Phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2), 11–20.

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS Computational Biology*, *7*(1), e1001052.

Barrett, J. (2007). Information processing in generalized probabilistic theories. *Physical Review A*, *75*(3), 032304.

Barrett, T. W., Manchak, J., & Weatherall, J. O. (2023). On automorphism criteria for comparing amounts of mathematical structure. *Synthese*, *201*(6), 191.

Bayne, T. (2005). Divided brains and unified phenomenology: a review essay on Michael Tye's consciousness and persons. *Philosophical Psychology*, *18*(4), 495–512.

Bayne, T. (2010). *The Unity of Consciousness*. Oxford University Press.

Bayne, T. (2018). On the axiomatic foundations of the Integrated Information Theory of consciousness. *Neuroscience of Consciousness*, *8*(1).

Bayne, T., & Chalmers, D. J. (2003). What is the Unity of Consciousness? In A. Cleeremans (Ed.), *The Unity of Consciousness*. Oxford University Press.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., … others (2024). Tests for consciousness in humans and beyond. *Trends in cognitive sciences*.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological review*, *75*(2), 127.

Beebee, H., Hitchcock, C., Menzies, P. C., & Menzies, P. (2009). *The Oxford Handbook of Causation*. Oxford Handbooks.

Bell, J. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, *1*(3), 195.

Bell, J. (1990). Against 'measurement'. *Physics world*, *3*(8), 33.

Benzmüller, C., & Lomfeld, B. (2020). Reasonable machines: A research manifesto. In *KI 2020: Advances in Artificial Intelligence* (pp. 251–258).

Bisiach, E., Luzzatti, C., & Perani, D. (1979). Unilateral neglect, representational schema and consciousness. *Brain*, *102*(3), 609–618.

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

Block, N. (1980). Troubles with functionalism. In *The Language and Thought Series* (pp. 268–306). Harvard University Press.

Block, N. (1990). Inverted earth. *Philosophical perspectives*, *4*, 53–79.

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

Block, N. (2004). Qualia. In R. L. Gregory (Ed.), *Oxford Companion to the Mind*. Oxford University Press.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer

science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences*, *119*(21), e2115934119.

Blum, L., & Blum, M. (2023). A theoretical computer science perspective on consciousness and artificial general intelligence. *Engineering*, *25*, 12–16.

Blum, M., & Blum, L. (2021). A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, *8*(01), 1–42.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., … others (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., … Tsuchiya, N. (2013). Consciousness in humans and non-human animals: recent advances and future directions. *Frontiers in psychology*, *4*, 625.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Bost rom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57–69). Chapman and Hall/CRC.

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155.

Bronfman, Z., Ginsburg, S., & Jablonka, E. (2021). When will robots be sentient? *Journal of Artificial Intelligence and Consciousness*, *8*(02), 183–203.

Brown, R. (in press). *Consciousness as Representing One's Mind: The Higher-Order Approach to Consciousness Explained*. Oxford University Press.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in cognitive sciences*, *23*(9), 754–768.

Brüntrup, G. (2011). Panpsychism and Structural Realism. In M. Blamauer (Ed.), *The Mental as Fundamental: New Perspectives on Panpsychism* (pp. 15–35). Ontos.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … others (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Capucci, M., Gavranović, B., Hedges, J., & Rischel, E. F. (2021). Towards foundations of categorical cybernetics. *arXiv preprint arXiv:2105.06332*.

Carroll, S. M. (2018). Beyond falsifiability: Normal science in a multiverse. *Why Trust a Theory?*, 300.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., … others (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, *5*(198), 198ra105–198ra105.

Casa rotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., … others (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of neurology*, *80*(5), 718–729.

Cerullo, M. A. (2015). The problem with Phi: a critique of Integrated Information Theory. *PLOS Computational Biology*, *11*(9).

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.
Chalmers, D. J. (1996). *The Conscious Mind: In search of a fundamental theory*. Oxford University Press.
Chalmers, D. J. (2004). How can we construct a science of consciousness? In *The Cognitive Neurosciences III* (pp. 1111–1119). Boston, MA: MIT Press.
Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*.
Chalmers, D. J. (2022). Carnap's Second Aufbau and David Lewis's Aufbau. In *Perspectives on the Philosophy of David K. Lewis*. Oxford University Press.
Chalmers, D. J. (2023a). Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.
Chalmers, D. J. (2023b). *Phenomenal Structuralism*. Talk presented at the Structuralism in Consciousness Studies Workshop at the Charité Berlin.
Chalmers, D. J., & McQueen, K. J. (2022). Consciousness and the Collapse of the Wave Function. In S. Gao (Ed.), *Consciousness and Quantum Mechanics*. Oxford University Press.
Chang, A. Y., Biehl, M., Yu, Y., & Kanai, R. (2020). Information closure theory of consciousness. *Frontiers in Psychology*, *11*.
Chiribella, G., D'Ariano, G. M., & Perinotti, P. (2010). Probabilistic theories with purification. *Physical Review A*, *81*(6), 062348.
Cho, K., & Jacobs, B. (2019). Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, *29*(7), 938–971.
Church, A. (1936). An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, *58*(2), 345–363.
Clancey, W. J. (1993). The strange, familiar, and forgotten: An anatomy of consciousness. *Artificial Intelligence*, *60*(2), 313–356.
Clark, A. (1993). *Sensory qualities*. Clarendon Library of Logic and Philosophy.
Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.
Clark, A. (2000). *A theory of sentience*. Clarendon press.
Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, *116*(12), 645–662.
Clark, S., Coecke, B., & Sadrzadeh, M. (2008). A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)* (pp. 133–140).
Cleeremans, A., & Frith, C. (2003). *The Unity of Consciousness*. Oxford University Press.
Cleeremans, A., & Tallon-Baudry, C. (2022). Consciousness matters: phenomenal experience has functional value. *Neuroscience of Consciousness*, *2022*(1), niac007.
Cliff, N. (1992). Abstract Measurement Theory and the Revolution that Never Happened. *Psychological Science*, *3*(3), 186–190.
Coecke, B. (2006). Introducing categories to the practicing physicist. In *What is category theory* (pp. 45–74).
Coecke, B. (2014). Terminality implies non-signalling. *arXiv preprint arXiv:1405.3681*.

Coecke, B., & Kissinger, A. (2017). *Picturing quantum processes*. Cambridge University Press.

Coecke, B., & Paquette, E. O. (2010). Categories for the practising physicist. In *New structures for physics* (pp. 173–286). Springer.

Coecke, B., Pavlovic, D., & Vicary, J. (2013). A new description of orthogonal bases. *Mathematical Structures in Computer Science*, *23*(3), 555–567.

Coecke, B., & Spekkens, R. W. (2012). Picturing classical and quantum Bayesian inference. *Synthese*, *186*, 651–696.

Coninx, S. (2022). A multidimensional phenomenal space for pain: structure, primitiveness, and utility. *Phenomenology and the Cognitive Sciences*, *21*(1), 223–243.

Cowey, A., & Stoerig, P. (1991). The neurobiology of blindsight. *Trends in Neurosciences*, *14*(4), 140–145.

CRediT taxonomy. (2021). CRediT taxonomy. In *JATS for Reuse*. JATS4R.

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

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences* (Vol. 2, pp. 263–275).

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

Deane, G. (2021). Consciousness in active inference: Deep self-models, other minds, and the challenge of psychedelic-induced ego-dissolution. *Neuroscience of Consciousness*, *2021*(2), niab024.

Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. *The cognitive neurosciences*, *3*, 1145–58.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227.

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications. *Characterizing consciousness: From cognition to the clinic?*, 55–84.

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

Dennett, D. C. (1988). Quining qualia. In A. Marcel & E. Bisiach (Eds.), *Consciousness in Modern Science*. Oxford University Press.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.

Dennett, D. C. (1995). *Darwin's Dangerous Idea – Evolution and the meanings of life*. Penguin Books.

De Santis, D. (2011). Phenomenological kaleidoscope: Remarks on the Husserlian method of eidetic variation. *New Yearbook for Phenomenology and Phenomenological Philosophy*, *11*, 16–41.

Devinsky, O., Farah, M. J., & Barr, W. B. (2008). Visual agnosia. *Handbook of Clinical Neurology*, *88*, 417–427.

De Vries, B., & Friston, K. J. (2017). A factor graph description of deep temporal active inference. *Frontiers in computational neuroscience*, *11*, 95.

Di Lavore, E., & Román, M. (2023). Evidential Decision Theory via Partial Markov Categories. *arXiv preprint arXiv:2301.12989*.

Dizadji-Bahmani, F., Frigg, R., & Hartmann, S. (2010). Who's afraid of Nagelian reduction? *Erkenntnis*, *73*(3), 393–412.

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and cognition*, *72*, 49–59.

Dolkega, K., & Dewhurst, J. E. (2020). Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*.

Edelman, G. M. (1989). *The remembered present: a biological theory of consciousness*. Basic Books.

Ehresmann, A. C. (2012). MENS: from neurons to higher mental processes up to consciousness. In *Integral Biomathics* (pp. 29–30). Springer.

Encyclopedia of Mathematics. (2013). Signed measure. In *Encyclopedia of Mathematics*. The European Mathematical Society.

Farrell, B. A. (1950). Experience. *Mind*, *59*(234), 170–198.

Fechner, G. (1860). *Elements of psychophysics. Vol. I.* New York.

Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and cognition*, *20*(3), 807–827.

Ferguson, A., Meyers, C., Bartlett, R., Banister, H., Bartlett, F., Brown, W., ... others (1940). Quantitative estimates of sensory events. Report of the British Association for the Advancement of Science. *The Advancement of Science*, *2*, 331–349.

Filk, T. (2016). It is the theory which decides what we can observe (Einstein). In *Contextuality from Quantum Physics to Psychology* (pp. 77–92). Singapore: World Scientific.

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*.

Fink, S. B., Wiese, W., & Windt, J. M. (2018). *Philosophical and Ethical Aspects of a Science of Consciousness and the Self*. Frontiers in Psychology.

Finster, F., & Kleiner, J. (2015). Causal fermion systems as a candidate for a unified physical theory. In *Journal of Physics: Conference Series* (Vol. 626).

Fong, B. (2013). Causal theories: A categorical perspective on Bayesian networks. *arXiv preprint arXiv:1301.6201*.

Fong, B., Spivak, D., & Tuyéras, R. (2019). Backprop as functor: A compositional perspective on supervised learning. In *2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)* (pp. 1–13).

Fortier-Davy, M., & Millière, R. (2020). The multi-dimensional approach to drug-induced states: A commentary on Bayne and Carter's "dimensions of consciousness and the psychedelic state". *Neuroscience of Consciousness*, *2020*(1), niaa004.

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

Frankish, K., & Sklutová, K. (2022). Illusionism and its place in contemporary philosophy of mind. *Human Affairs*, *32*(3), 300–310.

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

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5–19.

Frigg, R., & Votsis, I. (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European journal for philosophy of science*, *1*, 227–276.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, *29*(1), 1–49.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of physiology-Paris*, *100*(1-3), 70–87.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS one*, *4*(7), e6421.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network neuroscience*, *1*(4), 381–414.

Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., … others (2022). Designing Ecosystems of Intelligence from First Principles. *arXiv preprint arXiv:2212.01354*.

Frith, C., Perry, R., & Lumer, E. (1999). The neural correlates of conscious experience: An experimental framework. *Trends in cognitive sciences*, *3*(3), 105–114.

Fritz, T. (2020). A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, *370*, 107239.

Fritz, T., & Klingler, A. (2023). The d-separation criterion in categorical probability. *J. Mach. Learn. Res*, *24*(46), 1–49.

Ganesh, N. (2020). C-wars: the unfolding argument strikes back–a reply to 'falsification & consciousness'. *arXiv preprint arXiv:2006.13664*.

Gert, J. (2017). Quality spaces: Mental and physical. *Philosophical Psychology*, *30*(5), 525–544.

Ghani, N., Hedges, J., Winschel, V., & Zahn, P. (2018). Compositional game theory. In *Proceedings of the 33rd annual ACM/IEEE symposium on logic in computer science* (pp. 472–481).

Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1396*(1), 5.

Gödel, K. (1934). On undecidable propositions of formal mathematical systems. In *Collected Works* (Vol. 1). Oxford University Press, 1986.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press.

Gosseries, O., Di, H., Laureys, S., & Boly, M. (2014). Measuring consciousness in severely damaged brains. *Annual Review of Neuroscience*, *37*, 457–478.

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, 500.

Grindrod, P. (2018). On human consciousness: A mathematical perspective. *Network neuroscience*, *2*(1), 23–40.

Halmos, P. R. (1974). *Measure theory*. Springer.

Hanson, J. R., & Walker, S. I. (2019). Integrated Information Theory and Isomorphic Feed-Forward Philosophical Zombies. *Entropy*, *21*(11), 1073.

Hanson, J. R., & Walker, S. I. (2021). Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness*, *2021*(2), niab014.

Hanson, J. R., & Walker, S. I. (2023). On the non-uniqueness problem in Integrated Information Theory. *Neuroscience of Consciousness*, *2023*(1), niad014.

Hardy, L. (2017). Proposal to use humans to switch settings in a Bell experiment. *arXiv preprint arXiv:1705.04620*.

Harlan, R. M. (1984). Towards a computational phenomenology. *Man and World*, *17*(3), 261–277.

Hartmann, S., & Sprenger, J. (2012). The future of philosophy of science: introduction. *European Journal for Philosophy of Science*, *2*(2), 157–159.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, *21*(12), 1160.

Haun, A. M., Oizumi, M., Kovach, C. K., Kawasaki, H., Oya, H., Howard, M. A., … Tsuchiya, N. (2016). Contents of consciousness investigated as integrated information in direct human brain recordings. *bioRxiv*, 039032.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in cognitive sciences*, *13*(5), 194–202.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Heisenberg, W. (1971). *Physics and Beyond*. London: Allen & Unwin.

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In *Scientific explanation, space, and time.* University of Minnesota Press, Minneapolis.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of science*, *15*(2), 135–175.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*.

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.

Hobson, J. A., & Friston, K. J. (2014). Consciousness, dreams, and inference: The Cartesian theatre revisited. *Journal of Consciousness Studies*, *21*(1-2), 6–32.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology*, 577.

Hoffman, D. D., Prakash, C., & Prentner, R. (2023). Fusions of Consciousness. *Entropy*, *25*(1), 129.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 96.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209–223.

Hohwy, J., & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Holland, O. (2003). *Machine consciousness*. Imprint Academic.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Humphrey, N. (2023). *Sentience: The Invention of Consciousness*. MIT Press.

Husserl, E. (1936). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy*. Northwestern University Press, 1970.

Husserl, E. (1939). *Experience and Judgment*. Northwestern University Press, 1973.

Husserl, E. (1960). *Cartesian meditations: An introduction to phenomenology*. Springer.

Hutter, M. (2003). *A gentle introduction to the universal algorithmic agent AIXI*. Springer.

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

Irvine, E. (2012). *Consciousness as a scientific concept: A philosophy of science perspective*. Springer.

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

Irvine, E., & Sprevak, M. (2020). Eliminativism about consciousness. In U. Kriegel (Ed.), *Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press.

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, *83*(5), 291–295.

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

Jackson, F. (1998). Epiphenomenal qualia. In *Consciousness and emotion in cognitive science* (pp. 197–206). Routledge.

Jacobs, B. (2019). The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journal of Artificial Intelligence Research*, *65*, 783–806.

Jacobs, B., Kissinger, A., & Zanasi, F. (2019). Causal inference by string diagram surgery. In *Foundations of Software Science and Computation Structures* (pp. 313–329).

Ji, X., Elmoznino, E., Deane, G., Constant, A., Dumas, G., Lajoie, G., … Bengio, Y. (2024). Sources of richness and ineffability for phenomenally conscious states. *Neuroscience of Consciousness*, *2024*(1), niae001.

Josephs, E. L., Hebart, M. N., & Konkle, T. (2023). Dimensions underlying human understanding of the reachable world. *Cognition*, *234*, 105368.

Joshi, K. (1983). *Introduction to General Topology*. Wiley Eastern.

Joshi, K. D. (1989). *Foundations of Discrete Mathematics*. New Age International.

Jost, J. (2015). *Mathematical concepts*. Springer.

Juliani, A., Arulkumaran, K., Sasai, S., & Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*.

Juliani, A., Kanai, R., & Sasai, S. S. (2022). The Perceiver Architecture is a Functional Global Workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Kanai, R., & Fujisawa, I. (2024). Toward a universal theory of consciousness. *Neuroscience of Consciousness*, *2024*(1), niae022.

Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., & Oizumi, M. (2023). Comparing color similarity structures between humans and LLMs via unsupervised alignment. *arXiv preprint arXiv:2308.04381*.

Kawakita, G., Zeleznikow-Johnston, A. M., Takeda, K., Tsuchiya, N., & Oizumi, M. (2023). Is my "red" your "red"?: Unsupervised alignment of qualia structures via optimal transport. In *ICLR 2024 Workshop on Representational Alignment*.

Kent, A. (2018). Quanta and qualia. *Foundations of Physics*, *48*(9), 1021–1037.

Kent, A. (2020). Toy models of top down causation. *Entropy*, *22*(11), 1224.

Ketland, J. (2004). Empirical Adequacy and Ramsification. *British Journal for the Philosophy of Science*, *55*(2).

Kihlstrom, J. F. (2005). Dissociative disorders. *Annu. Rev. Clin. Psychol.*, *1*, 227–253.

Kim, J. (1996). *Philosophy of Mind*. Boulder: Westview Press.

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

Kissinger, A., & Uijlen, S. (2017). A categorical semantics for causal structure. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)* (pp. 1–12).

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In *Scientific explanation. Minnesota studies in the philosophy of science*. University of Minnesota Press, Minneapolis.

Kleene, S. C. (1936). λ-definability and recursiveness. *Duke Mathematical Journal*, *2*(2), 340 – 353.

Kleiner, J. (2020a). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition*, *85*.

Kleiner, J. (2020b). Mathematical models of consciousness. *Entropy*, *22*(6).

Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*.

Kleiner, J., & Hartmann, S. (2023). The Closure of the Physical, Consciousness and Scientific Practice. *arXiv preprint arXiv: 2110.03518*.

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1).

Kleiner, J., & Ludwig, T. (2024). What is a mathematical structure of conscious experience? *Synthese*, *203*(3), 89.

Kleiner, J., & Ludwig, T. (in press). The Case for Neurons: A No-Go Theorem for Consciousness on a Chip. *Neuroscience of Consciousness*.

Kleiner, J., & Tull, S. (2021). The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, *6*.

Klincewicz, M. (2011). Quality space model of temporal perception. In *Multidisciplinary aspects of time and time perception* (pp. 230–245). Springer.

Kob, L. (2023). Exploring the role of structuralist methodology in the neuroscience of consciousness: a defense and analysis. *Neuroscience of Consciousness*, *2023*(1), niad011.

Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness is Widespread But Can't be Computed*. Mit Press.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307.

Koch, R., & Murugan, J. (2012). Emergent spacetime. *Foundations of space and time: reflections on quantum gravity*, 164–184.

Kostic, D. (2012). The vagueness constraint and the quality space for pain. *Philosophical Psychology*, *25*(6), 929–939.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol. I-III*. Academic Press.

Kremnizer, K., & Ranchin, A. (2015). Integrated Information-induced quantum collapse. *Foundations of Physics*, *45*(8), 889–899.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Kuehni, R. (2010). Color spaces. *Scholarpedia*, *5*(3), 9606.

Kuehni, R. G., & Schwarz, A. (2008). *Color ordered: a survey of color systems from antiquity to the present*. Oxford University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lakatos, I. (1980). *The methodology of scientific research programmes: Volume 1: Philosophical papers* (Vol. 1). Cambridge University Press.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501.

Landau, L. D., & Lifshitz, E. M. (1980). *Statistical Physics Part 1*. Pergamon Press.

Lau, H., Michel, M., LeDoux, J. E., & Fleming, S. M. (2022). The mnemonic basis of subjective experience. *Nature Reviews Psychology*, *1*(8), 479–488.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences*, *15*(8), 365–373.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, A. Y. (2021). Modeling mental qualities. *Philosophical Review*, *130*(2), 263–298.

Lee, A. Y. (2022). Objective phenomenology. *Erkenntnis*, 1–20.

Lee, A. Y. (2023). Degrees of Consciousness. *Noûs*.

Leitgeb, H. (2013). Scientific philosophy, mathematical philosophy, and all that. *Metaphilosophy*, *44*(3), 267–275.

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

Lewis, C. I. (1929). *Mind and the World-Order: Outline of a theory of knowledge*. Courier Corporation.

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, *67*(13), 427–446.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. In *IFS: Conditionals, Belief, Decision, Chance and Time* (pp. 129–147). Springer.

Lewis, D. (1983). New work for a theory of universals. *Australasian journal of philosophy*, *61*(4), 343–377.

Lewis, D. (1986). *On the plurality of worlds*. Wiley-Blackwell.

List, C. (2019). Levels: descriptive, explanatory, and ontological. *Noûs*, *53*(4), 852–883.

Lorenz, R., & Tull, S. (2023). Causal models in string diagrams. *arXiv preprint arXiv:2304.07638*.

Luce, R. D., & Suppes, P. (2002). Representational Measurement Theory. In H. Pashler & J. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology*. John Wiley & Sons.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, *1*(1), 1–27.

Lyre, H. (2022). Neurophenomenal Structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1), niac012.

Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of consciousness*, *2021*(2), niab028.

Marshall, W., Gomez-Ramirez, J., & Tononi, G. (2016). Integrated information and state differentiation. *Frontiers in psychology*, *7*, 926.

Marshall, W., Kim, H., Walker, S. I., Tononi, G., & Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160358.

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

Mason, J. W. (2013). Consciousness and the structuring property of typical data. *Complexity*, *18*(3), 28–37.

Mason, J. W. (2016). Quasi-conscious multivariate systems. *Complexity*, *21*(S1), 125–147.

Mason, J. W. (2021). Model Unity and the Unity of Consciousness: Developments in Expected Float Entropy Minimisation. *Entropy*, *23*(11), 1444.

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, *309*(5744), 2228–2232.

Mather, J. A. (2008). Cephalopod consciousness: behavioural evidence. *Consciousness and cognition*, *17*(1), 37–48.

Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for Integrated Information Theory. *PLOS Computational Biology*, *14*(7), 1–21.

McDermott, D. (2007). Artificial intelligence and consciousness. *The Cambridge Handbook of Consciousness*, 117–150.

McQueen, K. J. (2019). Interpretation-Neutral Integrated Information Theory. *Journal of Consciousness Studies*, *26*(1-2), 76–106.

Mediano, P. A., Rosas, F., Carhart-Harris, R. L., Seth, A. K., & Barrett, A. B. (2019). Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv preprint arXiv:1909.02297*.

Mediano, P. A., Seth, A. K., & Barrett, A. B. (2019). Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy*, *21*(1), 17.

Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Jensen, H. J., Seth, A. K., Barrett, A. B., … Bor, D. (2022). Greater than the parts: a review of the information decomposition approach to causal emergence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *380*(2227), 20210246.

Melia, J., & Saatsi, J. (2006). Ramseyfication and theoretical content. *The British Journal for the Philosophy of Science*.

Metzinger, T. (1995). The problem of consciousness. In T. Metzinger (Ed.), *Conscious Experience* (pp. 3–37). Imprint Academic.

Metzinger, T. (2006). *Grundkurs Philosophie des Geistes, Band 1: Phänomenales Bewusstsein*. Paderborn.

Metzinger, T. (2007). *Philosophie des Bewusstseins*. Auditorum Netzwerk.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(01), 43–66.

Metzinger, T. (2024). *The elephant and the blind: the experience of pure consciousness: philosophy, science, and 500+ experiential reports*. MIT Press.

Michel, M. (2019). The mismeasure of consciousness: A problem of coordination for the perceptual awareness scale. *Philosophy of Science*, *86*(5), 1239–1249.

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), e1628.

Michel, M. (in press). The perceptual reality monitoring theory. In M. Herzog, A. Schurger, & A. Doerig (Eds.), *Scientific Theories of Consciousness: The Grand Tour*. Cambridge University Press.

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., … others (2019). Opportunities and challenges for a maturing science of consciousness. *Nature human behaviour*, *3*(2), 104–107.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

Mileti, J. (2022). *Modern Mathematical Logic*. Cambridge University Press.

Miller, M., Clark, A., & Schlicht, T. (2022). Predictive Processing and Consciousness. *Review of Philosophy and Psychology*, *13*(4), 797–808.

Mishra, P., & Dutt, N. D. (2005). *Functional verification of programmable embedded architectures: a top-down approach*. Springer Science & Business Media.

Mishra, P., Morad, R., Ziv, A., & Ray, S. (2017). Post-silicon validation in the SoC era: A tutorial introduction. *IEEE Design & Test*, *34*(3), 68–92.

Mitra, S., Seshia, S. A., & Nicolici, N. (2010). Post-silicon validation opportunities, challenges and recent advances. In *Proceedings of the 47th Design Automation Conference* (pp. 12–17).

Mollo, D. C., & Millière, R. (2023). The Vector Grounding Problem. *arXiv preprint arXiv:2304.01481*.

Montero, B., & Papineau, D. (2005). A defence of the via negativa argument for physicalism. *Analysis*, *65*(3), 233–237.

Moon, K., & Pae, H. (2018). Making Sense of Consciousness as Integrated Information: Evolution and Issues of IIT. *arXiv preprint arXiv:1807.02103*.

Mueller, M. P. (2017). Could the physical world be emergent instead of fundamental, and why should we ask? (Short version). *arXiv preprint arXiv:1712.01816*.

Myrvold, W. (2022). Philosophical Issues in Quantum Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 435–450.

Napolitani, M., Bodart, O., Canali, P., Seregni, F., Casali, A., Laureys, S., … Gosseries, O. (2014). Transcranial magnetic stimulation combined with high-density EEG in altered states of consciousness. *Brain injury*, *28*(9), 1180–1189.

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the Integrated Information Theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*, 1–18.

Newman, M. H. (1928). Mr. Russell's "Causal Theory of Perception". *Mind*, *37*(146), 137–148.

Nida-Rümelin, M. (2018). The experience property frame work: a misleading paradigm. *Synthese*, *195*, 3361–3387.

nLab. (2024a). *Faithful Functor.* (Revision 15)

nLab. (2024b). *Faithful Representation.* (Revision 8)

Northoff, G., Tsuchiya, N., & Saigo, H. (2019). Mathematics and the brain: A category theoretical approach to go beyond the neural correlates of consciousness. *Entropy*, *21*(12), 1234.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*(1), 127–148.

O'Connor, T. (2021a). Emergent Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

O'Connor, T. (2021b). Emergent Properties. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechan-

isms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5), 1–25.

Oxford English Dictionary. (1989). Methodology (n.). In *Second Edition.* Oxford University Press.

Ozawa, M. (1984). Quantum measuring processes of continuous observables. *Journal of Mathematical Physics*, *25*(1), 79–87.

Panangaden, P. (1998). Probabilistic relations. *School of Computer Science Research Reports-University of Birmingham CSR*, 59–74.

Papineau, D. (2009). The causal closure of the physical and naturalism. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind* (pp. 53–65). Oxford University Press.

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific reports*, *9*(1), 1889.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Pashler, H., & Wixted, J. (2004). *Stevens' Handbook of Experimental Psychology, Methodology in Experimental Psychology* (Vol. 4). John Wiley & Sons.

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

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

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peirce, C. S. (1866). Lowell Lecture. In M. H. Fisch (Ed.), *Writings of Charles S. Peirce: A Chronological Edition.* Indiana University Press.

Penrose, R. (1989). *The emperor's new mind*. Oxford University Press.

Peressini, A. (2013). Consciousness as Integrated Information a provisional philosophical critique. *Journal of Consciousness Studies*, *20*(1-2), 180–206.

Perrone, P. (2023). Markov categories and entropy. *IEEE Transactions on Information Theory*.

Petitot, J. (1999). Morphological eidetics for a phenomenology of perception. In J. Petitot, F. J. Varela, B. Pachoud, & J.-M. Roy (Eds.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 330–371). Stanford University Press.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford University Press.

Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press.

Piedeleu, R., & Zanasi, F. (2023). An Introduction to String Diagrams for Computer Scientists. *arXiv preprint arXiv:2305.08768*.

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Prentner, R. (2019). Consciousness and topologically structured phenomenal spaces. *Consciousness and Cognition*, *70*, 25–38.

Prentner, R. (2024a). Category theory in consciousness science: going beyond the correlational project. *Synthese*, *204*(2), 69.

Prentner, R. (2024b). Mathematized Phenomenology and the Science of Consciousness. *PsyArXiv Preprint*.

Psillos, S. (2005). *Scientific realism: How science tracks truth*. Routledge.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions Of Mind: A Symposium*. New York University Press.

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press. (Reprinted in (Putnam, 1975).)

Putnam, H. (1975). The Nature of Mental States. In *Mind, Language, and Reality: Philosophical Papers* (Vol. ii). Cambridge University Press.

Radnitzky, G. (1991). *Refined falsificationism meets the challenge from the relativist philosophy of science*. JSTOR.

Ramstead, M. J., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., ... others (2022). From generative models to generative passages: a computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4), 829–857.

Reardon, S. (2019). *Rival theories face off over brain's source of consciousness*. American Association for the Advancement of Science.

Renero, A. (2014). Consciousness and Mental Qualities for Auditory Sensations. *Journal of Consciousness Studies*, *21*(9-10), 179–204.

Resende, P. (2018). Quanta and Qualia. In *Proceedings of the Workshop on Combining Viewpoints in Quantum Theory, ICMS, Edinburgh, UK* (pp. 19–22).

Resende, P. (2022). Qualia as physical measurements: a mathematical model of qualia and pure concepts. *arXiv preprint arXiv:2203.10602*.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Robb, D., Heil, J., & Gibb, S. (2023). Mental Causation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Roelofs, L. (2016). The unity of consciousness, within subjects and between subjects. *Philosophical Studies*, *173*(12), 3199–3221.

Rosenthal, D. (1991). The independence of consciousness and sensory quality. *Philosophical issues*, *1*, 15–36.

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

Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, *20*, 368–393.

Rosenthal, D. (2015). Quality spaces and sensory modalities. In P. Coates & S. Coleman (Eds.), *Phenomenal qualities: sense, perception, and consciousness* (pp. 33–65). Oxford University Press.

Rosenthal, D. (2016). Quality spaces, relocation, and grain. In O'Shea (Ed.), *Sellars and his Legacy* (pp. 149–185). Oxford University Press.

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-

burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive neuroscience*, *1*(3), 165–175.
Rovelli, C. (2004). *Quantum gravity*. Cambridge University Press.
Rudin, W. (1976). *Principles of Mathematical Analysis* (Vol. 3). McGraw-Hill New York.
Rudin, W. (2006). *Real and complex analysis*. Tata McGraw-hill education.
Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, *428*, 106–131.
Russell, B. (2014). *The Autobiography of Bertrand Russell*. Routledge.
Safron, A. (2022). Integrated world modeling theory expanded: Implications for the future of consciousness. *Frontiers in Computational Neuroscience*, *16*, 642397.
Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: demystified and compared. *Neural computation*, *33*(3), 674–712.
Salamon, D. (2016). *Measure and integration*. European Mathematical Society.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Schäfer, A. M., & Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks* (pp. 632–640).
Schanda, J. (2007). CIE colorimetry. *Colorimetry: Understanding the CIE system*, *3*, 25–78.
Schlicht, T., & Dolega, K. (2021). You can't always get what you want. *Philosophy and the Mind Sciences*, *2*.
Seager, W. (2006). The 'Intrinsic Nature' argument for panpsychism. *Journal of Consciousness Studies*, *13*(10-11), 129–145.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.
Selby, J., & Coecke, B. (2017). Leaks: quantum, classical, intermediate and more. *Entropy*, *19*(4), 174.
Selinger, P. (2007). Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical computer science*, *170*, 139–163.
Selinger, P. (2011). A survey of graphical languages for monoidal categories. In *New structures for physics* (pp. 289–355). Springer.
Sergent, C., & Dehaene, S. (2004). Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *Journal of Physiology–Paris*, *98*(4–6), 374–384.
Seth, A. K. (2007). Models of consciousness. *Scholarpedia*, *2*(1), 1328.
Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, *5*(2), 97–118.
Seth, A. K. (2022). *Being you: A new science of consciousness*. Faber & Faber.
Seth, A. K., Barrett, A. B., & Barnett, L. (2011). Causal density and integrated information as measures of conscious level. *Philosophical Transactions of the Royal Society*

A: Mathematical, Physical and Engineering Sciences, *369*(1952), 3748–3767.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences*, *12*(8), 314–321.

Seth, A. K., & Hohwy, J. (2021). Predictive Processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, *12*(2), 89–90.

Sharma, G., Wu, W., & Dalal, E. N. (2004). The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations. *COLOR Research and Application*.

Shiebler, D., Gavranović, B., & Wilson, P. (2021). Category theory in machine learning. *arXiv preprint arXiv:2106.07032*.

Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, *79*(7), 357–381.

Shoemaker, S. (1991). Qualia and consciousness. *Mind*, *100*(4), 507–524.

Signorelli, C. M., Szczotka, J., & Prentner, R. (2021). Explanatory profiles of models of consciousness-towards a systematic classification. *Neuroscience of consciousness*, *2021*(2), niab021.

Signorelli, C. M., Wang, Q., & Coecke, B. (2021). Reasoning about conscious experience with axiomatic and graphical mathematics. *Consciousness and Cognition*, *95*, 103168.

Signorelli, C. M., Wang, Q., & Khan, I. (2021). A compositional model of consciousness based on consciousness-only. *Entropy*, *23*(3), 308.

Silva, L. (2023). Towards an Affective Quality Space. *Journal of Consciousness Studies*, *30*(7-8), 164–195.

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

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of mathematical psychology*, *107*, 102632.

Smithe, T. S. C. (2020). Bayesian Updates Compose Optically. *arXiv preprint arXiv:2006.01631*.

Smithe, T. S. C. (2021a). Compositional Active Inference I: Bayesian Lenses. Statistical Games. *arXiv preprint arXiv:2109.04461*.

Smithe, T. S. C. (2021b). Cyber kittens, or some first steps towards categorical cybernetics. *arXiv preprint arXiv:2101.10483*.

Smithe, T. S. C. (2022). Compositional Active Inference II: Polynomial Dynamics. Approximate Inference Doctrines. *arXiv preprint arXiv:2208.12173*.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and control*, *7*(1), 1–22.

Sprevak, M., & Irvine, E. (2020). Eliminativism about consciousness. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *Oxford Handbook of the Philosophy of Consciousness* (pp. 348–370). Oxford University Press.

Stanley, R. P. (1999). Qualia space. *Journal of Consciousness Studies*, *6*(1), 49–60.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stoerig, P. (2006). Blindsight, conscious vision, and the role of primary visual cortex. *Progress in Brain Research*, *155*, 217–234.

Strevens, M. (2006). Scientific explanation. In D. Borchert (Ed.), *Encyclopedia of Philosophy* (pp. 518–27). New York: Macmillan Reference USA.

Susskind, R. (2019). *Online courts and the future of justice*. Oxford University Press.

Tallon-Baudry, C. (2022). The topological space of subjective experience. *Trends in Cognitive Sciences*.

Tegmark, M. (2015). Consciousness as a state of matter. *Chaos, Solitons & Fractals*, *76*, 238–270.

Tegmark, M. (2016). Improved measures of integrated information. *PLoS Computational Biology*, *12*(11).

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

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

Tononi, G. (2005). Consciousness, information integration, and the brain. *Progress in brain research*, *150*, 109–126.

Tononi, G. (2008). Consciousness as Integrated Information: a provisional manifesto. *The Biological Bulletin*, *215*(3), 216–242.

Tononi, G. (2014). Why Scott should stare at a blank wall and reconsider (or, the conscious grid). *Shtetl-Optimized: The Blog of Scott Aaronson*.

Tononi, G. (2015). Integrated Information Theory. *Scholarpedia*, *10*(1), 4164.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450.

Tschantz, A., Millidge, B., Seth, A. K., & Buckley, C. L. (2020). Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*.

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition*, *79*, 102877.

Tsuchiya, N., Haun, A., Cohen, D., & Oizumi, M. (2016). Empirical tests of the Integrated Information Theory of consciousness. In *The return of consciousness: A new science on old questions* (pp. 349–374).

Tsuchiya, N., Phillips, S., & Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition*, *101*, 103319.

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034.

Tsuchiya, N., Saigo, H., & Phillips, S. (2023). An adjunction hypothesis between qualia and reports. *Frontiers in Psychology*, *13*, 1053977.

Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and Integrated Information Theory. *Neuroscience*

research, *107*, 1–7.

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

Tull, S., & Kleiner, J. (2021). Integrated Information in Process Theories: Towards Categorical IIT. *Journal of Cognitive Science*, *22*(2), 92–123.

Tull, S., Kleiner, J., & Smithe, T. S. C. (2023). Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy. *arXiv preprint arXiv:2308.00861*.

Turing, A. M. (1937a). Computability and λ-definability. *Journal of Symbolic Logic*, *2*(4), 153–163.

Turing, A. M. (1937b). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2*(1), 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*.

Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Stanford University.

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik (Mathematical Foundations of Quantum Mechanics)*. Julius Springer.

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

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

Wenzel, M., Han, S., Smith, E. H., Hoel, E., Greger, B., House, P. A., & Yuste, R. (2019). Reduced Repertoire of Cortical Microstates and Neuronal Ensembles in Medically Induced Loss of Consciousness. *Cell systems*, *8*(5), 467–474.

Wiese, W. (2018). What Is It Like to Experience a Third Man? The Phenomenological Bradley and How to Solve It. In W. Wiese (Ed.), *Experienced Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. The MIT Press.

Wiese, W. (2024). Artificial consciousness: a perspective from the free energy principle. *Philosophical Studies*, *181*(8), 1947–1970.

Wiese, W., & Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences*, *2*.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group.

Wile, B., Goss, J., & Roesner, W. (2005). *Comprehensive functional verification: The complete industry cycle*. Morgan Kaufmann.

Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan Publishing Co., New York.

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

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1–24.

Wu, T., & Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, *100*(3), 033311.

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour*, *6*(4), 593–604.

Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, *6*(3), 271–291.

Young, B. D., Keller, A., & Rosenthal, D. (2014). Quality-space theory in olfaction. *Frontiers in Psychology*, *5*, 1.

Zaidi, Q., Victor, J., McDermott, J., Geffen, M., Bensmaia, S., & Cleland, T. A. (2013). Perceptual spaces: mathematical structures to neural mechanisms. *Journal of Neuroscience*, *33*(45), 17597–17602.

Zanardi, P., Tomka, M., & Venuti, L. C. (2018). Towards quantum integrated information theory. *arXiv preprint arXiv:1806.01421*.

Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., & Tsuchiya, N. (2023). Are Color Experiences the Same across the Visual Field? *Journal of Cognitive Neuroscience*, *35*(4), 509–542.

Publications

Published:

Publication 1 (Chapter 2):

Kleiner, J., & Tull, S. (2021). The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, 6, 602973.

Publication 2 (Chapter 3):

Tull, S., & Kleiner, J. (2021). Integrated Information in Process Theories: Towards Categorical IIT. *Journal of Cognitive Science*, 22, 2, 92–123.

Publication 3 (Chapter 5):

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, 2021(1), niab001.

Publication 4 (Chapter 7):

Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, 119, 103653.

Publication 5 (Chapter 8):

Kleiner, J., & Ludwig, T. (2024). What is a mathematical structure of conscious experience?. *Synthese*, 203(3), 89.

Publication 6 (Chapter 10):

Kleiner, J., & Ludwig, T. (2024). The Case for Neurons: A No-Go Theorem for Consciousness on a Chip. Forthcoming in *Neuroscience of Consciousness*.

Under Review:

Preprint 7 (Chapter 4):

Tull, S., Kleiner, J., & Smithe, T. S. C. (2023). Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy.[9]

[9] Accepted to the 6th International Conference on Applied Category Theory (ACT2023).

Preprint 8 (Chapter 6):

Kleiner, J., & Hartmann, S. (2023). The Closure of the Physical, Consciousness and Scientific Practice.

Preprint 9 (Chapter 9):

Kleiner, J. (2024). The Newman Problem of Consciousness Science.

Preprint 10 (Chapter 11):

Kleiner, J. (2024). Consciousness qua Mortal Computation.

Preprint 11 (Chapters 1 and 12):

Kleiner, J. (2024). Mathematical Approaches in the Scientific Study of Consciousness.[10]

[10] Invited submission to Melloni, L. & Olcese, U. (Eds.). (Forthcoming). *The Scientific Study of Consciousness – Experimental and Theoretical Approaches*. Springer Nature.

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 04 June 2021
doi: 10.3389/fams.2020.602973

The Mathematical Structure of Integrated Information Theory

Johannes Kleiner[1]* *and Sean Tull*[2]*

[1]*Ludwig Maximilian University of Munich, Munich, Germany,* [2]*Cambridge Quantum Computing Limited, Cambridge, United Kingdom*

Integrated Information Theory is one of the leading models of consciousness. It aims to describe both the quality and quantity of the conscious experience of a physical system, such as the brain, in a particular state. In this contribution, we propound the mathematical structure of the theory, separating the essentials from auxiliary formal tools. We provide a definition of a generalized IIT which has IIT 3.0 of Tononi et al., as well as the Quantum IIT introduced by Zanardi et al. as special cases. This provides an axiomatic definition of the theory which may serve as the starting point for future formal investigations and as an introduction suitable for researchers with a formal background.

Keywords: Integrated Information Theory, experience spaces, mathematical consciousness science, IIT 3.0, IIT 3.x, generalized IIT

Edited by:
Heng Liu,
Guangxi University for Nationalities, China

Reviewed by:
Guangming Xue,
Guangxi University of Finance and Economics, China
Shumin Ha,
Shaanxi Normal University, China

***Correspondence:**
Johannes Kleiner
johannes.kleiner@lmu.de
Sean Tull
sean.tull@cambridgequantum.com

Specialty section:
This article was submitted to Dynamical Systems, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *07 September 2020*
Accepted: *23 December 2020*
Published: *04 June 2021*

Citation:
Kleiner J and Tull S (2021) The Mathematical Structure of Integrated Information Theory. Front. Appl. Math. Stat. 6:602973. doi: 10.3389/fams.2020.602973

1 INTRODUCTION

Integrated Information Theory (IIT), developed by Giulio Tononi and collaborators [5, 45–47], has emerged as one of the leading scientific theories of consciousness. At the heart of the latest version of the theory [19, 25, 26, 31, 40] is an algorithm which, based on the level of *integration* of the internal functional relationships of a physical system in a given state, aims to determine both the quality and quantity ('Φ value') of its conscious experience.

While promising in itself [12, 43], the mathematical formulation of the theory is not satisfying to date. The presentation in terms of examples and accompanying explanation veils the essential mathematical structure of the theory and impedes philosophical and scientific analysis. In addition, the current definition of the theory can only be applied to comparably simple classical physical systems [1], which is problematic if the theory is taken to be a fundamental theory of consciousness, and should eventually be reconciled with our present theories of physics.

To resolve these problems, we examine the essentials of the IIT algorithm and formally define a generalized notion of Integrated Information Theory. This notion captures the inherent mathematical structure of IIT and offers a rigorous mathematical definition of the theory which has 'classical' IIT 3.0 of Tononi et al. [25, 26, 31] as well as the more recently introduced *Quantum Integrated Information Theory* of Zanardi, Tomka and Venuti [50] as special cases. In addition, this generalization allows us to extend classical IIT, freeing it from a number of simplifying assumptions identified in [3]. Our results are summarised in **Figure 1**.

In the associated article [44] we show more generally how the main notions of IIT, including causation and integration, can be treated, and an IIT defined, starting from any suitable theory of physical systems and processes described in terms of category theory. Restricting to classical or quantum process then yields each of the above as special cases. This treatment makes IIT applicable to a large class of physical systems and helps overcome the current restrictions.

Our definition of IIT may serve as the starting point for further mathematical analysis of IIT, in particular if related to category theory [30, 49]. It also provides a simplification and mathematical clarification of the IIT algorithm which extends the technical analysis of the theory [1, 41, 42] and may contribute to its ongoing critical discussion [2, 4, 8, 23, 27, 28, 33]. The concise presentation of IIT in this article should also help to make IIT more easily accessible for mathematicians, physicists and other researchers with a strongly formal background.

This work is concerned with the most recent version of IIT as proposed in [25, 26, 31, 40] and similar papers quoted below. Thus our constructions recover the specific theory of consciousness referred to as IIT 3.0 or IIT 3.x, which we will call classical IIT in what follows. Earlier proposals by Tononi et al. that also aim to explicate the general idea of an essential connection between consciousness and integrated information constitute alternative theories of consciousness which we do not study here. A yet different approach would be to take the term 'Integrated Information Theory' to refer to the general idea of associating conscious experience with some pre-theoretic notion of integrated information, and to explore the different ways that this notion could be defined in formal terms [4, 27, 28, 37].

Relation to Other Work

This work develops a thorough mathematical perspective of one of the promising contemporary theories of consciousness. As such it is part of a number of recent contributions which seek to explore the role and prospects of mathematical theories of consciousness [11, 15, 18, 30, 49], to help overcome problems of existing models [17, 18, 34] and to eventually develop new proposals [6, 13, 16, 20, 22, 29, 39].

1.1 Structure of Article

We begin by introducing the necessary ingredients of a generalised Integrated Information Theory in **Sections 2–4**, namely physical systems, experience spaces and cause-effect repertoires. Our approach is *axiomatic* in that we state only the precise formal structure which is necessary to apply the IIT algorithm. We neither motivate nor criticize these structures as necessary or suitable to model consciousness. Our goal is simply to recover IIT 3.0. In **Section 5**, we introduce a simple formal tool which allows us to present the definition of the algorithm of an IIT in a concise form in **Sections 6** and **7**. Finally, in **Section 8**, we summarise the full definition of such a theory. The result is the definition of a generalized IIT. We call any application of this definition 'an IIT'.

Following this we give several examples including IIT 3.0 in **Section 9** and Quantum IIT in **Section 10**. In **Section 11** we discuss how our formulation allows one to extend classical IIT in several fundamental ways, before discussing further modifications to our approach and other future work in **Section 12**. Finally, the appendix includes a detailed explanation of how our generalization of IIT coincides with its usual presentation in the case of classical IIT.

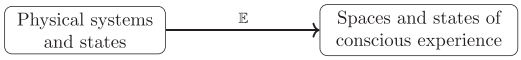


FIGURE 1 | An Integrated Information Theory specifies for every system in a particular state its conscious experience, described formally as an element of an experience space. In our formalization, this is a map **Sys** $\xrightarrow{\mathbb{E}}$ **Exp** from the system class **Sys** into a class **Exp** of experience spaces, which, first, sends each system S to its space of possible experiences $\mathbb{E}(S)$, and, second, sends each state $s \in \mathrm{St}(S)$ to the actual experience the system is having when in that space, $\mathrm{St}(S) \to \mathbb{E}(S)$, $s \mapsto \mathbb{E}(S,s)$. The definition of this map in terms of axiomatic descriptions of physical systems, experience spaces and further structure used in classical IIT is given in the first half of this paper.

2 SYSTEMS

The first step in defining an Integrated Information Theory (IIT) is to specify a class **Sys** of physical *systems* to be studied. Each element $S \in$ **Sys** is interpreted as a model of one particular physical system. In order to apply the IIT algorithm, it is only necessary that each element S come with the following features.

Definition 1. A *system class* **Sys** is a class each of whose elements S, called *systems*, come with the following data:

1. A set $\mathrm{St}(S)$ of *states*;
2. for every $s \in \mathrm{St}(S)$, a set $\mathrm{Sub}_s(S) \subset$ **Sys** of subsystems and for each $M \in \mathrm{Sub}_s(S)$ an induced state $s|_M \in \mathrm{St}(M)$;
3. a set $\mathbb{D}_S$ of *decompositions*, with a given *trivial decomposition* $1 \in \mathbb{D}_S$;
4. for each $z \in \mathbb{D}_S$ a corresponding *cut system* $S^z \in$ **Sys** and for each state $s \in \mathrm{St}(S)$ a corresponding *cut state* $s^z \in \mathrm{St}(S^z)$. Moreover, we require that **Sys** contains a distinguished *empty system*, denoted I, and that $I \in \mathrm{Sub}(S)$ for all S. For the IIT algorithm to work, we need to assume furthermore that the number of subsystems remains the same under cuts or changes of states, i.e. that we have bijections $\mathrm{Sub}_s(S) \simeq \mathrm{Sub}_{s'}(S)$ for all $s, s' \in \mathrm{St}(S)$ and $\mathrm{Sub}_s(S) \simeq \mathrm{Sub}_{s^z}(S^z)$ for all $z \in \mathbb{D}_S$.

   Note that taking a subsystem of a system S requires specifying a state s of S. An example class of systems is illustrated in **Figure 2**. In this article we will assume that each set $\mathrm{Sub}_s(S)$ is finite, discussing the extension to the infinite case in **Section 12**. We will give examples of system classes and for all following definitions in **Sections 9** and **10**.

3 EXPERIENCE

An IIT aims to specify for each system in a particular state its *conscious experience*. As such, it will require a mathematical model of such experiences. Examining classical IIT, we find the following basic features of the final experiential states it describes which are needed for its algorithm.

Firstly, each experience e should crucially come with an *intensity*, given by a number $||e||$ in the non-negative reals $\mathbb{R}^+$ (including zero). This intensity will finally correspond to the overall intensity of experience, usually denoted by Φ. Next, in order to compare experiences, we require a notion of *distance* $d(e,e')$ between any

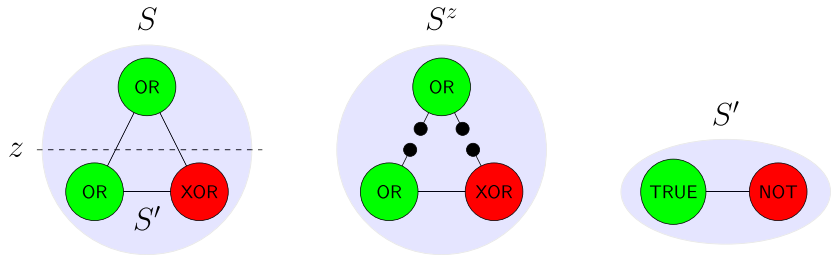


FIGURE 2 | As an example of Definition 1 similar to IIT 3.0, consider simple systems given by sets of nodes (or 'elements'), with a state assigning each node the state 'on' (depicted green) or 'off' (red). Each system comes with a time evolution shown by labelling each node with how its state in the next time-step depends on the states of the others. Decompositions of a system S correspond to binary partition of the nodes, such as z above. The cut system S^z is given by modifying the time evolution of S so that the two halves do not interact; in this case all connections between the halves are replaced by sources of noise which send 'on' or 'off' with equal likelihood, depicted as black dots above. Given a current state s of S, any subset of the nodes (such as those below the dotted line) determines a subsystem S', with time evolution obtained from that of S by fixing the nodes in $S \setminus S'$ (here, the upper node) to be in the state specified by s. Note that while in this example any subsystem (subset of S) determines a decomposition (partition of S) we do not require such a relationship in general.

pair of experiences e,e'. Finally, the algorithm will require us to be able to *rescale* any given experience e to have any given intensity. Mathematically, this is most easily encoded by letting us multiply any experience e by any number $r \in \mathbb{R}^+$. In summary, a minimal model of experience in a generalized IIT is the following.

Definition 2. An *experience space* is a set E with:

1. An *intensity* function $||.|| : E \to \mathbb{R}^+$
2. A *distance* function $d : E \times E \to \mathbb{R}^+$
3. A *scalar multiplication* $\mathbb{R}^+ \times E \to E$, denoted $(r, e) \mapsto r \cdot e$, satisfying

$$\|r \cdot e\| = r \cdot \|e\| \quad r \cdot (s \cdot e) = (rs) \cdot e \quad 1 \cdot e = e$$

for all $e \in E$ and $r, s \in \mathbb{R}^+$.

We remark that this same axiomatisation will apply both to the full space of experiences of a system, as well as to the spaces describing components of the experiences ('concepts' and 'proto-experiences' defined in later sections). We note that the distance function does not necessarily have to satisfy the axioms of a metric. While this and further natural axioms such as $d(r \cdot e, r \cdot f) = r \cdot d(e, f)$ might hold, they are not necessary for the IIT algorithm.

The above definition is very general, and in any specific application of IIT, the experiences may come with further mathematical structure. The following example includes the experience spaces used in classical IIT.

Example 3. Any metric space (X, d) may be extended to an experience space $\overline{X} := X \times \mathbb{R}^+$ in various ways. E.g., one can define $||(x, r)|| = r$, $r \cdot (x, s) = (x, rs)$ and define the distance as

$$d\big((x, r), (y, s)\big) = r\, d(x, y). \tag{1}$$

This is the definition used in classical IIT (cf. **Section 9** and **Appendix A**).

An important operation on experience spaces is taking their *product*.

Definition 4. For experience spaces E and F, we define the product to be the space $E \times F$ with distance

$$d\big((e, f), (e', f')\big) = d(e, e') + d(f, f'), \tag{2}$$

intensity $\|(e, f)\| = \max\{\|e\|, \|f\|\}$ and scalar multiplication $r \cdot (e, f) = (r \cdot e, r \cdot f)$. This generalizes to any finite product $\prod_{i \in I} E_i$ of experience spaces.

4 REPERTOIRES

In order to define the experience space and individual experiences of a system S, an IIT utilizes basic building blocks called 'repertoires', which we will now define. Next to the specification of a system class, this is the essential data necessary for the IIT algorithm to be applied.

Each repertoire describes a way of 'decomposing' experiences, in the following sense. Let D denote any set with a distinguished element 1, for example the set $\mathbb{D}_S$ of decompositions of a system S, where the distinguished element is the trivial decomposition $1 \in \mathbb{D}_S$.

Definition 5. Let e be an element of an experience space E. A *decomposition of e over D* is a mapping $\overline{e} : D \to E$ with $\overline{e}(1) = e$.

In more detail, a repertoire specifies a proto-experience for every pair of subsystems and describes how this experience changes if the subsystems are decomposed. This allows one to assess how integrated the system is with respect to a particular repertoire. Two repertoires are necessary for the IIT algorithm to be applied, together called the cause-effect repertoire.

For subsystems $M, P \in \mathrm{Sub}_s(S)$, define $\mathbb{D}_{M,P} := \mathbb{D}_M \times \mathbb{D}_P$. This set describes the decomposition of both subsystems simultaneously. It has a distinguished element $1 = (1_M, 1_P)$.

Definition 6. A *cause-effect repertoire* at S is given by a choice of experience space $\mathbb{PE}(S)$, called the space of *proto-experiences*, and for each $s \in \mathrm{St}(S)$ and $M, P \in \mathrm{Sub}_s(S)$, a pair of elements

$$\mathrm{caus}_s(M,P), \mathrm{eff}_s(M,P) \in \mathbb{PE}(S) \tag{3}$$

and for each of them a decomposition over $\mathbb{D}_{M,P}$.

Examples of cause-effect repertoires will be given in **Sections 9** and **10**. A general definition in terms of process theories is given in [44]. For the IIT algorithm, a cause-effect repertoire needs to be specified for every system S, as in the following definition.

Definition 7. A *cause-effect structure* is a specification of a cause-effect repertoire for every $S \in \mathbf{Sys}$ such that

$$\mathbb{PE}(S) = \mathbb{PE}(S^z) \quad \text{for all} \quad z \in \mathbb{D}_S. \tag{4}$$

The names 'cause' and 'effect' highlight that the definitions of $\mathrm{caus}_s(M,P)$ and $\mathrm{eff}_s(M,P)$ in classical and Quantum IIT describe the causal dynamics of the system. They are intended to capture the manner in which the 'current' state s of the system, when restricted to M, constrains the 'previous' or 'next' state of P, respectively.

5 INTEGRATION

We have now introduced all of the data required to define an IIT; namely, a system class along with a cause-effect structure. From this, we will give an algorithm aiming to specify the conscious experience of a system. Before proceeding to do so, we introduce a conceptual short-cut which allows the algorithm to be stated in a concise form. This captures the core ingredient of an IIT, namely the computation of how integrated an entity is.

Definition 8. Let E be an experience space and e an element with a decomposition over some set D. The *integration level* of e relative to this decomposition is

$$\phi(e) := \min_{1 \neq z \in D} d(e, \bar{e}(z)). \tag{5}$$

Here, d denotes the distance function of E, and the minimum is taken over all elements of D besides 1. The *integration scaling* of e is then the element of E defined by

$$\iota(e) := \phi(e) \cdot \hat{e}, \tag{6}$$

where $\hat{e}$ denotes the *normalization* of e, defined as

$$\hat{e} := \begin{cases} \dfrac{1}{\|e\|} \cdot e & \text{if } \|e\| \neq 0 \\ e & \text{if } \|e\| = 0. \end{cases}$$

Finally, the *integration scaling* of a pair e_1, e_2 of such elements is the pair

$$\iota(e_1, e_2) := \left(\phi \cdot \hat{e_1}, \phi \cdot \hat{e_2}\right) \tag{7}$$

where $\phi := \min(\phi(e_1), \phi(e_2))$ is the minimum of their integration levels.

We will also need to consider indexed collections of decomposable elements. Let S be a system in a state $s \in \mathrm{St}(S)$ and assume that for every $M \in \mathrm{Sub}_s(S)$ an element e_M of some experience space E_M with a decomposition over some D_M is given. We call $(e_M)_{M \in \mathrm{Sub}_s(S)}$ a *collection* of decomposable elements, and denote it as $(e_M)_M$.

Definition 9. The *core* of the collection $(e_M)_M$ is the subsystem $C \in \mathrm{Sub}_s(S)$ for which $\phi(e_C)$ is maximal.[1] The *core integration scaling* of the collection is $\iota(e_C)$. The *core integration scaling* of a pair of collections $((e_M)_M, (f_M)_M)$ is $\iota(e_C, f_D)$, where C, D are the cores of $(e_M)_M$ and $(f_M)_M$, respectively.

6 CONSTRUCTIONS: MECHANISM LEVEL

Let $S \in \mathbf{Sys}$ be a physical system whose experience in a state $s \in \mathrm{St}(S)$ is to be determined. The first level of the algorithm involves fixing some subsystem $M \in \mathrm{Sub}_s(S)$, referred to as a 'mechanism', and associating to it an object called its 'concept' which belongs to the *concept space*

$$\mathbb{C}(S) := \mathbb{PE}(S) \times \mathbb{PE}(S). \tag{8}$$

For every choice of $P \in \mathrm{Sub}_s(S)$, called a 'purview', the repertoire values $\mathrm{caus}_s(M,P)$ and $\mathrm{eff}_s(M,P)$ are elements of $\mathbb{PE}(S)$ with given decompositions over $\mathbb{D}_{M,P}$. Fixing M, they provide elements with decompositions over Sub(S) given by

$$\begin{aligned} \mathrm{caus}_s(M) &:= (\mathrm{caus}_s(M,P))_{P \in \mathrm{Sub}(S)} \\ \mathrm{eff}_s(M) &:= (\mathrm{eff}_s(M,P))_{P \in \mathrm{Sub}(S)}. \end{aligned} \tag{9}$$

The *concept* of M is then defined as the core integration scaling of this pair of collections,

$$\mathbb{C}_{S,s}(M) := \text{Core integration scaling of } \left(\mathrm{caus}_s(M), \mathrm{eff}_s(M)\right). \tag{10}$$

It is an element of $\mathbb{C}(S)$. Unraveling our definitions, the concept thus consists of the values of the cause and effect repertoires at their respective 'core' purviews P^c, P^e, i.e. those which make them 'most integrated'. These values $\mathrm{caus}(M, P^c)$ and $\mathrm{eff}(M, P^e)$ are then each rescaled to have intensity given by the minima of their two integration levels.

7 CONSTRUCTIONS: SYSTEM LEVEL

The second level of the algorithm specifies the experience of system S in state s. To this end, all concepts of a system are collected to form its *Q-shape*, defined as

$$\mathbb{Q}_s(S) := \left(\mathbb{C}_{S,s}(M)\right)_{M \in \mathrm{Sub}_s(S)}. \tag{11}$$

This Is an Element of the Space

$$\mathbb{E}(S) = \mathbb{C}(S)^{n(S)}, \tag{12}$$

where $n(S) := |\mathrm{Sub}_s(S)|$, which is finite and independent of the state s according to our assumptions. We can also define a Q-shape for any cut of S. Let $z \in \mathbb{D}_S$ be a decomposition, S^z the corresponding cut system and s^z be the corresponding cut state. We define

[1] If the maximum does not exist, we define the core to be the empty system I.

$$\mathbb{Q}_s(S^z) := \left(\mathbb{C}_{S^z,s^z}(M)\right)_{M\in \mathrm{Sub}_{s^z}(S^z)}. \tag{13}$$

Because of **Eq. 4**, and since the number of subsystems remains the same when cutting, $\mathbb{Q}_s(S^z)$ is also an element of $\mathbb{E}(S)$. This gives a map

$$\begin{aligned}\overline{\mathbb{Q}}_{S,s} : \mathbb{D}_S &\to \mathbb{E}(S)\\ z &\mapsto \mathbb{Q}_s(S^z)\end{aligned}$$

which is a decomposition of $\mathbb{Q}_s(S)$ over $\mathbb{D}_S$. Considering this map for every subsystem of S gives a collection of decompositions defined as

$$\mathbb{Q}(S,s) := \left(\overline{\mathbb{Q}}_{M,s|_M}\right)_{M\in \mathrm{Sub}_s(S)}$$

This is the system level-object of relevance and is what specifies the experience of a system according to IIT.

Definition 10. The *experience* of system S in the state $s \in \mathrm{St}(S)$ is

$$\mathbb{E}(S,s) := \text{Core integration scaling of } \mathbb{Q}(S,s). \tag{14}$$

The definition implies that $\mathbb{E}(S,s) \in \mathbb{E}(M)$, where $M \in \mathrm{Sub}_s(S)$ is the core of the collection $\mathbb{Q}(S,s)$, called the *major complex*. It describes which part of system S is actually conscious. In most cases there will be a natural embedding $\mathbb{E}(M) \to \mathbb{E}(S)$ for a subsystem M of S, allowing us to view $\mathbb{E}(S,s)$ as an element of $\mathbb{E}(S)$ itself. Assuming this embedding to exist allows us to define an Integrated Information Theory concisely in the next section.

8 INTEGRATED INFORMATION THEORIES

We can now summarize all that we have said about IITs.

Definition 11. An *Integrated Information Theory* is determined as follows. The *data* of the theory is a system class **Sys** along with a cause-effect structure. The theory then gives a mapping

$$\mathbf{Sys} \xrightarrow{\mathbb{E}} \mathbf{Exp} \tag{15}$$

into the class **Exp** of all experience spaces, sending each system S to its space of experiences $\mathbb{E}(S)$ defined in **Eq. 12**, and a mapping

$$\begin{aligned}\mathrm{St}(S) &\to \mathbb{E}(S)\\ s &\mapsto \mathbb{E}(S,s)\end{aligned} \tag{16}$$

which determines the experience of the system when in a state s, defined in **Eq. 14**.
The *quantity* of the system's experience is given by

$$\Phi(S,s) := \|\mathbb{E}(S,s)\|,$$

and the quality of the system's experience is given by the normalized experience $\widehat{\mathbb{E}}(S,s)$. The experience is "located" in the core of the collection $\mathbb{Q}(S,s)$, called *the major complex*, which is a subsystem of S.

In the next sections we specify the data of several example IITs.

9 CLASSICAL IIT

In this section we show how IIT 3.0 [25, 26, 31, 48] fits in into the framework developed here. A detailed explanation of how our earlier algorithm fits with the usual presentation of IIT is given in **Appendix A**. In [44] we give an alternative categorical presentation of the theory.

9.1 Systems

We first describe the system class underlying classical IIT. Physical systems S are considered to be built up of several components $S_1, \ldots, S_n$, called *elements*. Each element S_i comes with a finite set of states $\mathrm{St}(S_i)$, equipped with a metric. A state of S is given by specifying a state of each element, so that

$$\mathrm{St}(S) = \mathrm{St}(S_1) \times \cdots \times \mathrm{St}(S_n). \tag{17}$$

We define a metric d on $\mathrm{St}(S)$ by summing over the metrics of the element state spaces $\mathrm{St}(S_i)$ and denote the collection of probability distributions over $\mathrm{St}(S)$ by $\mathcal{P}(S)$. Note that we may view $\mathrm{St}(S)$ as a subset of $\mathcal{P}(S)$ by identifying any $s \in \mathrm{St}(S)$ with its Dirac distribution $\delta_s \in \mathcal{P}(S)$, which is why we abbreviate δ_s by s occasionally in what follows.

Additionally, each system comes with a probabilistic (discrete) *time evolution operator* or *transition probability matrix*, sending each $s \in \mathrm{St}(S)$ to a probabilistic state $T(s) \in \mathcal{P}(S)$. Equivalently it may be described as a convex-linear map

$$T : \mathcal{P}(S) \to \mathcal{P}(S). \tag{18}$$

Furthermore, the evolution T is required to satisfy a property called *conditional independence*, which we define shortly.

The class **Sys** consists of all possible tuples $S = (\{S_i\}_{i=1}^n, T)$ of this kind, with the trivial system I having only a single element with a single state and trivial time evolution.

9.2 Conditioning and Marginalizing

In what follows, we will need to consider two operations on the map T. Let M be any subset of the elements of a system and $M^\perp$ its complement. We again denote by $\mathrm{St}(M)$ the Cartesian product of the states of all elements in M, and by $\mathcal{P}(M)$ the probability distributions on $\mathrm{St}(M)$. For any $p \in \mathcal{P}(M)$, we define the *conditioning* [26] of T on p as the map

$$\begin{aligned}T|p\rangle : \mathcal{P}(M^\perp) &\to \mathcal{P}(S)\\ p' &\mapsto T(p \cdot p')\end{aligned} \tag{19}$$

where $p \cdot p'$ denotes the multiplication of these probability distributions to give a probability distribution over S. Next, we define *marginalisation over M* as the map

$$\langle M| : \mathcal{P}(S) \to \mathcal{P}(M^\perp) \tag{20}$$

such that for each $p \in \mathcal{P}(S)$ and $m_2 \in \mathrm{St}(M^\perp)$ we have

$$\langle M|p(m_2) = \sum_{m_1 \in \mathrm{St}(M)} p(m_1, m_2). \tag{21}$$

In particular for any map T as above we call $\langle M|T$ the *marginal* of T over M and we write $T_i := \langle S_i^\perp \big| T$ for each $i = 1, \ldots, n$. Conditional independence of T may now be defined as the requirement that

$$T(p) = \prod_{i=1}^{n} T_i(p) \quad \text{for all } p \in \mathcal{P}(S),$$

where the right-hand side is again a probability distribution over $\mathrm{St}(S)$.

9.3 Subsystems, Decompositions and Cuts

Let a system S in a state $s \in \mathrm{St}(S)$ be given. The subsystems are characterized by subsets of the elements that constitute S. For any subset $M = \{S_1, \ldots, S_m\}$ of the elements of S, the corresponding subsystem is also denoted M and $\mathrm{St}(M)$ is again given by the product of the $\mathrm{St}(S_i)$, with time evolution

$$T_M := \langle M^{\perp} | T | s_{M^{\perp}} \rangle, \tag{22}$$

where $s_{M^{\perp}}$ is the restriction of the state s to $\mathrm{St}(M^{\perp})$ and $|s_{M^{\perp}}\rangle$ denotes the conditioning on the Dirac distribution $\delta_{s_{M^{\perp}}}$.

The decomposition set $\mathbb{D}_S$ of a system S consists of all partitions of the set N of elements of S into two disjoint sets M and $M^{\perp}$. We denote such a partition by $z = (M, M^{\perp})$. The trivial decomposition 1 is the pair $(N, \varnothing)$.

For any decomposition $(M, M^{\perp})$ the corresponding cut system $S^{(M,M^{\perp})}$ is the same as S but with a new time evolution $T^{(M,M^{\perp})}$. Using conditional independence, it may be defined for each $i = 1, \ldots, n$ as

$$T_i^{(M,M^{\perp})} := \begin{cases} T_i & i \in M^{\perp} \\ T_i | \omega_{M^{\perp}} \rangle \langle M^{\perp} | & i \in M \end{cases}, \tag{23}$$

where $\omega_M \in \mathcal{P}(M)$ denotes the uniform distribution on $\mathrm{St}(M)$. This is interpreted in the graph depiction as removing all those edges from the graph whose source is in $M^{\perp}$ and whose target is in M. The corresponding input of the target element is replaced by noise, i.e. the uniform probability distribution over the source element.

9.4 Proto-Experiences

For each system S, the first Wasserstein metric (or 'Earth Mover's Distance') makes $\mathcal{P}(S)$ a metric space $(\mathcal{P}(S), d)$. The space of proto-experiences of classical IIT is

$$\mathbb{PE}(S) := \overline{\mathcal{P}(S)}, \tag{24}$$

where $\overline{\mathcal{P}(S)}$ is defined in Example 3. Thus elements of $\mathbb{PE}(S)$ are of the form (p, r) for some $p \in \mathcal{P}(S)$ and $r \in \mathbb{R}^+$, with distance function, intensity and scalar multiplication as defined in the example.

9.5 Repertoires

It remains to define the cause-effect repertoires. Fixing a state s of S, the first step will be to define maps caus_s' and eff_s' which send any choice of $(M, P) \in \mathrm{Sub}(S) \times \mathrm{Sub}(S)$ to an element of $\mathcal{P}(P)$. These should describe the way in which the current state of M constrains that of P in the next or previous time-steps. We begin with the effect repertoire. For a single element purview P_i we define

$$\mathrm{eff}_s'(M, P_i) := \langle P_i^{\perp} \big| T | \omega_{M^{\perp}} \rangle (s_M), \tag{25}$$

where s_M denotes (the Dirac distribution of) the restriction of the state s to M. While it is natural to use the same definition for arbitrary purviews, IIT 3.0 in fact uses another definition based on consideration of 'virtual elements' [25, 26, 48], which also makes calculations more efficient (Supplementary Material S1 of [26]). For general purviews P, this definition is

$$\mathrm{eff}_s'(M, P) = \prod_{P_i \in P} \mathrm{eff}_s'(M, P_i), \tag{26}$$

taking the product over all elements P_i in the purview P. Next, for the cause repertoire, for a single element mechanism M_i and each $\tilde{s} \in \mathrm{St}(P)$, we define

$$\mathrm{caus}_s'(M_i, P)[\tilde{s}] = \lambda \langle M_i^{\perp} \big| T | \omega_{P^{\perp}} \rangle (\delta_{\tilde{s}}) \big[s_{M_i} \big], \tag{27}$$

where λ is the unique normalisation scalar making $\mathrm{caus}_s'(M_i, P)$ a valid element of $\mathcal{P}(P)$. Here, for clarity, we have indicated evaluation of probability distributions at particular states by square brackets. If the time evolution operator has an inverse T^{-1}, this cause repertoire could be defined similarly to (25) by $\mathrm{caus}_s'(M_i, P) = \langle P^{\perp} \big| T^{-1} \big| \omega_{M_i^{\perp}} \rangle (s_{M_i})$, but classical IIT does not utilize this definition.

For General Mechanisms M, we Then Define

$$\mathrm{caus}_s'(M, P) = \kappa \prod_{M_i \in M} \mathrm{caus}_s'(M_i, P) \tag{28}$$

where the product is over all elements M_i in M and where $\kappa \in \mathbb{R}^+$ is again a normalisation constant. We may at last now define

$$\mathrm{caus}_s(M, P) := \mathrm{caus}_s'(M, P) \cdot \mathrm{caus}_s'(\varnothing, P^{\perp})$$

$$\mathrm{eff}_s(M, P) := \mathrm{eff}_s'(M, P) \cdot \mathrm{eff}_s'(\varnothing, P^{\perp}), \tag{29}$$

with intensity 1 when viewed as elements of $\mathbb{PE}(S)$. Here, the dot indicates again the multiplication of probability distributions and $\varnothing$ denotes the empty mechanism.

The distributions $\mathrm{caus}_s'(\varnothing, P^{\perp})$ and $\mathrm{eff}_s'(\varnothing, P^{\perp})$ are called the *unconstrained cause and effect repertoires* over $P^{\perp}$.

Remark 12. It is in fact possible for the right-hand side of **Eq. 28** to be equal to 0 for all $\tilde{s}$ for some $M_i \in M$. In this case we set $\mathrm{caus}_s'(M, P) = (\omega_S, 0)$ in $\mathbb{PE}(S)$.

Finally we must specify the decompositions of these elements over $\mathbb{D}_{M,P}$. For any partitions $z_M = (M_1, M_2)$ of M and $z_P = (P_1, P_2)$ of P, we define

$$\overline{\mathrm{caus}_s}(M,P)(z_M, z_P) := \mathrm{caus}_s'(M_1, P_1) \cdot \mathrm{caus}_s'(M_2, P_2) \cdot \mathrm{caus}_s'(\varnothing, P^{\perp})$$

$$\overline{\mathrm{eff}_s}(M,P)(z_M, z_P) := \mathrm{eff}_s'(M_1, P_1) \cdot \mathrm{eff}_s'(M_2 P_2) \cdot \mathrm{eff}_s'(\varnothing, P^{\perp}), \tag{30}$$

where we have abused notation by equating each subset M_1 and M_2 of nodes with their induced subsystems of S via the state s.

This concludes all data necessary to define classical IIT. If the generalized definition of **Section 8** is applied to this data, it yields precisely classical IIT 3.0 defined by Tononi et al. In Appendix A, we explain in detail how our definition of IIT, equipped with this data, maps to the usual presentation of the theory.

10 QUANTUM IIT

In this section, we consider Quantum IIT defined in [50]. This is also a special case of the definition in terms of process theories we give in [44].

10.1 Systems

Similar to classical IIT, in Quantum IIT systems are conceived as consisting of elements $\mathcal{H}_1, \ldots, \mathcal{H}_n$. Here, each element $\mathcal{H}_i$ is described by a finite dimensional Hilbert space and the state space of system S is defined in terms of the element Hilbert spaces as

$$\mathrm{St}(S) = \mathcal{S}(\mathcal{H}_S) \quad \text{with} \quad \mathcal{H}_S = \bigotimes_{i=1}^{n} \mathcal{H}_i,$$

where $\mathcal{S}(\mathcal{H}_S) \subset L(\mathcal{H}_S)$ describes the positive semidefinite Hermitian operators of unit trace on $\mathcal{H}_S$, i.e. density matrices. The time evolution of the system is again given by a time evolution operator, which here is assumed to be a trace preserving (and in [50] typically unital) completely positive map

$$\mathcal{T} : L(\mathcal{H}_S) \to L(\mathcal{H}_S).$$

10.2 Subsystems, Decompositions and Cuts

Subsystems are again defined to consist of subsets M of the elements of the system, with corresponding Hilbert space $\mathcal{H}_M := \bigotimes_{i \in M} \mathcal{H}_i$. The time-evolution $\mathcal{T}_M : L(\mathcal{H}_M) \to L(\mathcal{H}_M)$ is defined as

$$\mathcal{T}_M(\rho) = \mathrm{tr}_{M^\perp}\left(\mathcal{T}\left(\mathrm{tr}_{M^\perp}(s) \otimes \rho\right)\right),$$

where $s \in \mathcal{S}(\mathcal{H}_S)$ and $\mathrm{tr}_{M^\perp}$ denotes the trace over the Hilbert space $\mathcal{H}_{M^\perp}$.

Decompositions are also defined via partitions $z = (D, D^\perp) \in \mathbb{D}_S$ of the set of elements N into two disjoint subsets D and $D^\perp$ whose union is N. For any such decomposition, the cut system $S^{(D,D^\perp)}$ is defined to have the same set of states but time evolution

$$\mathcal{T}^{(D,D^\perp)}(s) = \mathcal{T}\left(\mathrm{tr}_{D^\perp}(s) \otimes \omega_{D^\perp}\right),$$

where $\omega_{D^\perp}$ is the maximally mixed state on $\mathcal{H}_{D^\perp}$, i.e. $\omega_{D^\perp} = \frac{1}{\dim(\mathcal{H}_{D^\perp})} 1_{\mathcal{H}_{D^\perp}}$.

10.3 Proto-Experiences

For any $\rho, \sigma \in \mathcal{S}(\mathcal{H}_S)$, the trace distance defined as

$$d(\rho, \sigma) = \frac{1}{2} \mathrm{tr}_S\left(\sqrt{(\rho - \sigma)^2}\right)$$

turns $(S(\mathcal{H}_S), d)$ into a metric space. The space of proto-experiences is defined based on this metric space as described in Example 3,

$$\mathbb{PE}(S) := \overline{S(\mathcal{H}_S)}.$$

10.4 Repertoires

We finally come to the definition of the cause-effect repertoire. Unlike classical IIT, the definition in [50] does not consider virtual elements. Let a system S in state $s \in \mathrm{St}(S)$ be given. As in **Section 9.5**, we utilize maps caus_s' and eff_s' which here map subsystems M and P to $\mathrm{St}(P)$. They are defined as

$$\mathrm{eff}_s'(M, P) = tr_{P^\perp} \mathcal{T}\left(tr_{M^\perp}(s) \otimes \omega_{M^\perp}\right)$$

$$\mathrm{caus}_s'(M, P) = tr_{P^\perp} \mathcal{T}^\dagger\left(tr_{M^\perp}(s) \otimes \omega_{M^\perp}\right),$$

where $\mathcal{T}^\dagger$ is the Hermitian adjoint of $\mathcal{T}$. We then define

$$\mathrm{caus}_s(M, P) := \mathrm{caus}_s'(M, P) \otimes \mathrm{caus}_s'(\varnothing, P^\perp)$$

$$\mathrm{eff}(M, P) := \mathrm{eff}_s'(M, P) \otimes \mathrm{eff}_s'(\varnothing, P^\perp),$$

each with intensity 1, where $\varnothing$ again denotes the empty mechanism. Similarly, decompositions of these elements over $\mathbb{D}_{M,P}$ are defined as

$$\overline{\mathrm{caus}_s}(M,P)(z_M, z_P) := \mathrm{caus}_s'(M_1, P_1) \otimes \mathrm{caus}_s'(M_2, P_2) \otimes \mathrm{caus}_s'(\varnothing, P^\perp)$$

$$\overline{\mathrm{eff}_s}(M,P)(z_M, z_P) := \mathrm{eff}_s'(M_1, P_1) \otimes \mathrm{eff}_s'(M_2, P_2) \otimes \mathrm{eff}_s'(\varnothing, P^\perp),$$

again with intensity 1, where $z_M = (M_1, M_2) \in \mathbb{D}_M$ and $z_P = (P_1, P_2) \in \mathbb{D}_P$.

11 EXTENSIONS OF CLASSICAL IIT

The physical systems to which IIT 3.0 may be applied are limited in a number of ways: they must have a discrete time-evolution, satisfy Markovian dynamics and exhibit a discrete set of states [3]. Since many physical systems do not satisfy these requirements, if IIT is to be taken as a fundamental theory about reality, it must be extended to overcome these limitations.

In this section, we show how IIT can be redefined to cope with continuous time, non-Markovian dynamics and non-compact state spaces, by a redefinition of the maps **Eqs. 26** and **28** and, in the case of non-compact state spaces, a slightly different choice of **Eq. 24**, while leaving all of the remaining structure as it is. While we do not think that our particular definitions are satisfying as a general definition of IIT, these results show that the disentanglement of the essential mathematical structure of IIT from auxiliary tools (the particular definition of cause-effect repertoires used to date) can help to overcome fundamental mathematical or conceptual problems.

In **Section 11.3**, we also explain which solution to the problem of non-canonical metrics is suggested by our formalism.

11.1 Discrete Time and Markovian Dynamics

In order to avoid the requirement of a discrete time and Markovian dynamics, instead of working with the time evolution operator **Eq. 18**, we define the cause- and effect repertoires in reference to a given trajectory of a physical state $s \in \mathrm{St}(S)$. The resulting definitions can be applied independently of whether trajectories are being determined by Markovian dynamics in a particular application, or not.

Let $t \in \mathcal{I}$ denote the time parameter of a physical system. If time is discrete, $\mathcal{I}$ is an ordered set. If time is continuous, $\mathcal{I}$ is an interval of reals. For simplicity, we assume $0 \in \mathcal{I}$. In the deterministic case, a trajectory of a state $s \in \mathrm{St}(S)$ is simply a curve in $\mathrm{St}(S)$, which we denote by $(s(t))_{t \in \mathcal{I}}$ with $s(0) = s$. For probabilistic systems (such as neural networks with a probabilistic update rule), it is a curve of probability distributions $\mathcal{P}(S)$, which we denote by $(p(t))_{t \in \mathcal{I}}$, with $p(0)$

equal to the Dirac distribution δ_s. The latter case includes the former, again via Dirac distributions.

In what follows, we utilize the fact that in physics, state spaces are defined such that the dynamical laws of a system allow to determine the trajectory of each state. Thus for every $s \in \mathrm{St}(S)$, there is a trajectory $(p_s(t))_{t\in\mathcal{I}}$ which describes the time evolution of s.

The idea behind the following is to define, for every $M, P \in \mathrm{Sub}(S)$, a trajectory $p_s^{(P,M)}(t)$ in $\mathcal{P}(P)$ which quantifies how much the state of the purview P at time t is being constrained by imposing the state s at time $t = 0$ on the mechanism M. This gives an alternative definition of the maps (26) and (28), while the rest of classical IIT can be applied as before.

Let now $M, P \in \mathrm{Sub}(S)$ and $s \in \mathrm{St}(S)$ be given. We first consider the time evolution of the state $(s_M, \nu) \in \mathrm{St}(S)$, where s_M denotes the restriction of s to $\mathrm{St}(M)$ as before and where $\nu \in \mathrm{St}(M^\perp)$ is an arbitrary state of $M^\perp$. We denote the time evolution of this state by $p_{(s_M,\nu)}(t) \in \mathcal{P}(S)$. Marginalizing this distribution over $P^\perp$ gives a distribution on the states of P, which we denote as $p^P_{(s_M,\nu)}(t) \in \mathcal{P}(P)$. Finally, we average over ν using the uniform distribution $\omega_{M^\perp}$. Because state spaces are finite in classical IIT, this averaging can be defined pointwise for every $w \in \mathrm{St}(P)$ by

$$p_s^{(P,M)}(t)(w) := \kappa \sum_{\nu \in St(M^\perp)} p^P_{(s_M,\nu)}(t)(w)\,\omega_{M^\perp}(\nu), \tag{31}$$

where κ is the unique normalization constant which ensures that $p_s^{(P,M)}(t) \in \mathcal{P}(P)$.

The probability distribution $p_s^{(P,M)}(t) \in \mathcal{P}(P)$ describes how much the state of the purview P at time t is being constrained by imposing the state s on M at time $t = 0$ as desired. Thus, for every $t \in \mathcal{I}$, we have obtained a mapping of two subsystems M, P to an element $p_s^{(P,M)}(t)$ of $\mathcal{P}(P)$ which has the same interpretation as the map **Eq. 26** considered in classical IIT. If deemed necessary, virtual elements could be introduced just as in **Eqs 27** and **29**.

So far, our construction can be applied for any time $t \in T$. It remains to fix this freedom in the choice of time. For the discrete case, the obvious choice is to define **Eqs 27** and **29** in terms of neighboring time-steps. For the continuous case, several choices exist. E.g., one could consider the positive and negative semi-derivatives of $p_s^{(P,M)}(t)$ at $t = 0$, in case they exist, or add an arbitrary but fixed time scale Δ to define the cause-effect repertoires in terms of $p_s^{(P,M)}(t_0 \pm \Delta)$. However, the most reasonable choice is in our eyes to work with limits, in case they exist, by defining

$$\mathrm{eff}'_s(M,P) := \prod_{P_i \in P} \lim_{t\to\infty} p_s^{(P_i,M)}(t) \tag{32}$$

to replace **Eq. 27** and

$$\mathrm{caus}'_s(M,P) := \kappa \prod_{M_i \in M} \lim_{t\to-\infty} p_s^{(P,M_i)}(t) \tag{33}$$

to replace **Eq. 29**. The remainder of the definitions of classical IIT can then be applied as before.

11.2 Discrete Set of States

The problem with applying the definitions of classical IIT to systems with continuous state spaces (e.g., neuron membrane potentials [3]) is that in certain cases, uniform probability distributions do not exist. E.g., if the state space of a system S consists of the positive real numbers $\mathbb{R}^+$, no uniform distribution can be defined which has a finite total volume, so that no uniform *probability* distribution ω_S exists.

It is important to note that this problem is less universal than one might think. E.g., if the state space of the system is a closed and bounded subset of $\mathbb{R}^+$, e.g. an interval $[a, b] \subset \mathbb{R}^+$, a uniform probability distribution can be defined using measure theory, which is in fact the natural mathematical language for probabilities and random variables. Nevertheless, the observation in [3] is correct that if a system has a non-compact continuous state space, ω_S might not exist, which can be considered a problem w.r.t. the above-mentioned working hypothesis.

This problem can be resolved for all well-understood physical systems by replacing the uniform probability distribution ω_S by some other mathematical entity which allows to define a notion of averaging states. For all relevant classical systems with non-compact state spaces (whether continuous or not), there exists a canonical uniform measure μ_S which allows to define the cause-effect repertoires similar to the last section, as we now explain. Examples for this canonical uniform measure are the Lebesgue measure for subsets of $\mathbb{R}^n$ [35], or the Haar measure for locally compact topological groups [36] such as Lie-groups.

In what follows, we explain how the construction of the last section needs to be modified in order to be applied to this case.

In all relevant classical physical theories, $\mathrm{St}(S)$ is a metric space in which every probability measure is a Radon measure, in particular locally finite, and where a canonical locally finite uniform measure μ_S exists. We define $\mathcal{P}_1(S)$ to be the space of probability measures whose first moment is finite. For these, the first Wasserstein metric (or 'Earth Mover's Distance') W_1 exists, so that $(\mathcal{P}_1(S), W_1)$ is a metric space.

As before, the dynamical laws of the physical systems determine for every state $s \in \mathrm{St}(S)$ a time evolution $p_s(t)$, which here is an element of $\mathcal{P}_1(S)$. Integration of this probability measure over $\mathrm{St}(P^\perp)$ yields the marginal probability measure $p_s^P(t)$. As in the last section, we may consider these probability measures for the state $(s_M, \nu) \in \mathrm{St}(S)$, where $\nu \in \mathrm{St}(M^\perp)$. Since μ_S is not normalizable, we cannot define $p_s^{(P,M)}(t)$ as in **(32)**, for the result might be infinite.

Using the fact that μ_S is locally finite, we may, however, define a somewhat weaker equivalent. To this end, we note that for every state $s_{M^\perp}$, the local finiteness of $\mu_{M^\perp}$ implies that there is a neighborhood $N_{s,M^\perp}$ in $\mathrm{St}(M^\perp)$ for which $\mu_{M^\perp}(N_{s,M^\perp})$ is finite. We choose a sufficiently large neighborhood which satisfies this condition. Assuming $p^P_{(s_M,\nu)}(t)$ to be a measurable function in ν, for every A in the σ-algebra of $\mathrm{St}(M^\perp)$, we can thus define

$$p_s^{(P,M)}(t)(A) := \kappa \int_{N_{s,M^\perp}} p^P_{(s_M,\nu)}(t)(A)\, d\mu_{M^\perp}(\nu), \tag{34}$$

which is a finite quantity. The $p_s^{(P,M)}(t)$ so defined is non-negative, vanishes for $A = \varnothing$ and satisfies countable additivity. Hence it is a measure on $\mathrm{St}(P)$ as desired, but might not be normalizable.

All that remains for this to give a cause-effect repertoire as in the last section, is to make sure that any measure (normalized or not) is an element of $\mathbb{PE}(S)$. The theory is flexible enough to do this by setting $d(\mu,\nu) = |\mu - \nu|(\mathrm{St}(P))$ if either μ or ν is not in $\mathcal{P}_1(S)$, and $W_1(\mu,\nu)$ otherwise. Here, $|\mu - \nu|$ denotes the total variation of the signed measure $\mu - \nu$, and $|\mu - \nu|(\mathrm{St}(P))$ is the volume thereof [10, 32]. While not a metric space any more, the tuple $(\mathcal{M}(S), d)$, with

$\mathcal{M}(S)$ denoting all measures on St(S), can still be turned into a space of proto-experiences as in Example 3. This gives

$$\mathbb{PE}(S) := \overline{\mathcal{M}(S)}$$

and finally allows one to construct cause-effect repertoires as in the last section.

11.3 Non-canonical Metrics

Another criticism of IIT's mathematical structure mentioned [3] is that the metrics used in IIT's algorithm are, to a certain extend, chosen arbitrarily. Different choices indeed imply different results of the algorithm, both concerning the quantity and quality of conscious experience, which can be considered problematic.

The resolution of this problem is, however, not so much a technical as a conceptual or philosophical task, for what is needed to resolve this issue is a justification of why a particular metric should be used. Various justifications are conceivable, e.g. identification of desired behavior of the algorithm when applied to simple systems. When considering our mathematical reconstruction of the theory, the following natural justification offers itself.

Implicit in our definition of the theory as a map from systems to experience spaces is the idea that the mathematical structure of experiences spaces (Definition 2) reflects the phenomenological structure of experience. This is so, most crucially, for the distance function d, which describes how similar two elements of experience spaces are. Since every element of an experience space corresponds to a conscious experience, it is naturally to demand that the similarly of the two mathematical objects should reflect the similarity of the experiences they describe. Put differently, the distance function d of an experience space should in fact mirror (or "model") the similarity of conscious experiences as experienced by an experiencing subject.

This suggests that the metrics d used in the IIT algorithm should, ultimately, be defined in terms of the phenomenological structure of similarity of conscious experiences. For the case of color qualia, this is in fact feasible [18, Example 3.18], [21, 38]. In general, the mathematical structure of experience spaces should be intimately tied to the phenomenology of experience, in our eyes.

12 SUMMARY AND OUTLOOK

In this article, we have propounded the mathematical structure of Integrated Information Theory. First, we have studied which exact structures the IIT algorithm uses in the mathematical description of physical systems, on the one hand, and in the mathematical description of conscious experience, on the other. Our findings are the basis of definitions of a physical system class **Sys** and a class **Exp** of experience spaces, and allowed us to view IIT as a map **Sys** $\to$ **Exp**.

Next, we needed to disentangle the essential mathematics of the theory from auxiliary formal tools used in the contemporary definition. To this end, we have introduced the precise notion of decomposition of elements of an experience space required by the IIT algorithm. The pivotal cause-effect repertoires are examples of decompositions so defined, which allowed us to view any particular choice, e.g. the one of 'classical' IIT developed by Tononi et al., or the one of 'quantum' IIT recently introduced by Zanardi et al. as data provided to a general IIT algorithm.

The formalization of cause-effect repertoires in terms of decompositions then led us to define the essential ingredients of IIT's algorithm concisely in terms of integration levels, integration scalings and cores. These definitions describe and unify recurrent mathematical operations in the contemporary presentation, and finally allowed to define IIT completely in terms of a few lines of definition.

Throughout the paper, we have taken great care to make sure our definitions reproduce exactly the contemporary version of IIT 3.0. The result of our work is a mathematically rigorous and general definition of Integrated Information Theory. This definition can be applied to any meaningful notion of systems and cause-effect repertoires, and we have shown that this allows one to overcome most of the mathematical problems of the contemporary definition identified to date in the literature.

We believe that our mathematical reconstruction of the theory can be the basis for refined mathematical and philosophical analysis of IIT. We also hope that this mathematisation may make the theory more amenable to study by mathematicians, physicists, computer scientists and other researchers with a strongly formal background.

12.1 Process Theories

Our generalization of IIT is axiomatic in the sense that we have only included those formal structures in the definition which are necessary for the IIT algorithm to be applied. This ensured that our reconstruction is as general as possible, while still true to IIT 3.0. As a result, several notions used in classical IIT, e.g., system decomposition, subsystems or causation, are merely defined abstractly at first, without any reference to the usual interpretation of these concepts in physics.

In the related article [44], we show that these concepts can be meaningfully defined in any suitable *process theory* of physics, formulated in the language of *symmetric monoidal categories*. This approach can describe both classical and Quantum IIT and yields a complete formulation of contemporary IIT in a categorical framework.

12.2 Further Development of IIT

IIT is constantly under development, with new and refined definitions being added every few years. We hope that our mathematical analysis of the theory might help to contribute to this development. For example, the working hypothesis that IIT is a fundamental theory, implies that technical problems of the theory need to be resolved. We have shown that our formalization allows one to address the technical problems mentioned in the literature. However, there are others which we have not addressed in this paper.

Most crucially, the IIT algorithm uses a series of maximalization and minimalization operations, unified in the notion of *core* subsystems in our formalization. In general, there is no guarantee that these operations lead to unique results, neither in classical nor Quantum IIT. Using different cores has major impact on the output of the algorithm, including the Φ value, which is a case of ill-definedness.[2]

[2]The problem of 'unique existence' has been studied extensively in category theory using *universal properties* and the notion of a *limit*. Rather than requiring that each $E \in \mathbb{E}$ come with a metric, it may be possible to alter the IIT algorithm into a well-defined categorical form involving limits to resolve this problem.

Furthermore, the contemporary definition of IIT as well as our formalization rely on there being a finite number of subsystems of each system, which might not be the case in reality. Our formalisation may be extendable to the infinite case by assuming that every system has a fixed but potentially infinite indexing set $\mathrm{Sub}(S)$, so that each $\mathrm{Sub}_s(S)$ is the image of a mapping $\mathrm{Sub}(S) \times \mathrm{St}(S) \to \mathbf{Sys}$, but we have not considered this in detail in this paper.

Finally, concerning more operational questions, it would be desirable to develop the connection to empirical measures such as the Perturbational Complexity Index (PCI) [7, 9] in more detail, as well as to define a controlled approximation of the theory whose calculation is less expensive. Both of these tasks may be achievable by substituting parts of our formalization with simpler mathematical structure.

On the conceptual side of things, it would be desirable to have a more proper understanding of how the mathematical structure of experiences spaces corresponds to the phenomenology of experience, both for the general definition used in our formalization—which comprises the minimal mathematical structure which is required for the IIT algorithm to be applied—and the specific definitions used in classical and Quantum IIT. In particular, it would be desirable to understand how it relates to the important notion of qualia, which is often asserted to have characteristic features such as ineffability, intrinsicality, non-contextuality, transparency or homogeneity [24]. For a first analysis toward this goal, cf [18]. A first proposal to add additional structure to IIT that accounts for relations between elements of consciousness in the case of spatial experiences was recently given in [14].

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JK and ST conceived the project together and wrote the article together.

ACKNOWLEDGMENTS

We would like to thank the organizers and participants of the *Workshop on Information Theory and Consciousness* at the Center for Mathematical Sciences of the University of Cambridge, of the *Modeling Consciousness Workshop* in Dorfgastein and of the *Models of Consciousness Conference* at the Mathematical Institute of the University of Oxford for discussions on this topic. Much of this work was carried out while Sean Tull was under the support of an EPSRC Doctoral Prize at the University of Oxford, from November 2018 to July 2019, and while Johannes Kleiner was under the support of postdoctoral funding at the Institute for Theoretical Physics of the Leibniz University of Hanover. We would like to thank both institutions.

REFERENCES

1. Barrett AB. An Integration of Integrated Information Theory with Fundamental Physics. *Front Psychol* (2014) 5:63. doi:10.3389/fpsyg.2014.00063
2. Tim B. On the Axiomatic Foundations of the Integrated Information Theory of Consciousness. *Neurosci Conscious* (2018). 2018(1) niy007. doi:10.1093/nc/niy007
3. Barrett AB, and Mediano PAM. The Phi Measure of Integrated Information Is Not Well-Defined for General Physical Systems. *J Conscious Stud* (2019). 21:133. doi:10.1021/acs.jpcb.6b05183.s001
4. Adam B, and Seth A. Practical Measures of Integrated Information for Time-Series Data. *Plos Comput Biol* (2011). 7(1):e1001052. doi:10.1371/journal.pcbi.1001052
5. Balduzzi D, and Tononi G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *Plos Comput Biol* (2008). 4(6):e1000091. doi:10.1371/journal.pcbi.1000091
6. Chang AYC, Biehl M, Yen Y, and Kanai R. Information Closure Theory of Consciousness. *Front Psychol* (2020). 11:121. doi:10.3389/fpsyg.2020.01504
7. Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, et al. Stratification of Unresponsive Patients by an Independently Validated Index of Brain Complexity. *Ann Neurol* (2016). 80(5):718–29. doi:10.1002/ana.24779
8. Boly MA. The Problem with Phi: a Critique of Integrated Information Theory. *Plos Comput Biol* (2015). 11(9):e1004286. doi:10.1371/journal.pcbi.1004286
9. Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, et al. A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Sci Translat Med* (2015). 5:198ra105. doi:10.1126/scitranslmed.3006294
10. Halmos PR. *Measure Theory* Berlin: Springer (1974).
11. Hardy. L. Proposal to Use Humans to Switch Settings in a Bell Experiment (2017). arXiv preprint arXiv:1705.04620.
12. Haun AM, Oizumi M, Kovach CK, Kawasaki H, Oya H, Howard MA, et al. Contents of consciousness investigated as integrated information in direct human brain recordings. (2016). bioRxiv.
13. Hoffman DD, and Prakash C. Objects of Consciousness. *Front Psychol* (2014). 5:577. doi:10.3389/fpsyg.2014.00577
14. Haun A, and Tononi G. Why Does Space Feel the Way it Does? towards a Principled Account of Spatial Experience. *Entropy* (2019). 21(12):1160. doi:10.3390/e21121160
15. Kent A. Quanta and Qualia. *Found Phys* (2018). 48(9):1021–37. doi:10.1007/s10701-018-0193-9
16. Kent A. Toy Models of Top Down Causation. (2019). arXiv preprint arXiv:1909.12739.
17. Kleiner J, and Hoel E. Falsification and Consciousness. *Neurosci Consciousness* (2021). 2021(1):niab001. doi:10.1093/nc/niab001
18. Kleiner J. Mathematical Models of Consciousness. *Entropy* (2020). 22(6):609. doi:10.3390/e22060609
19. Koch C, Massimini M, Boly M, and Tononi G. Neural Correlates of Consciousness: Progress and Problems. *Nat Rev Neurosci* (2016). 17(5):307, 21. doi:10.1038/nrn.2016.22
20. Kremnizer K, and Ranchin A. Integrated Information-Induced Quantum Collapse. *Found Phys* (2015). 45(8):889–99. doi:10.1007/s10701-015-9905-6
21. Kuehni R. Color Spaces. *Scholarpedia* (2010). 5(3):9606. doi:10.4249/scholarpedia.9606
22. Mason JWD. Quasi-conscious Multivariate Systems. *Complexity* (2016). 21(S1):125–47. doi:10.1002/cplx.21720
23. Kelvin J. Interpretation-neutral Integrated Information Theory. *J Conscious Stud* (2019). 26(1-2):76–106. doi:10.1007/978-1-4419-9707-4_13
24. Metzinger T. *Grundkurs Philosophie des Geistes, Band 1*. Berlin: Phänomenales Bewusstsein (2006).

25. Marshall W, Gomez-Ramirez J, and Tononi G. Integrated Information and State Differentiation. *Front Psychol* (2016). 7:926. doi:10.3389/fpsyg.2016.00926
26. William G, Mayner P, Marshall W, Albantakis L, Findlay G, Marchman R, et al. PyPhi: A Toolbox for Integrated Information Theory. *Plos Comput Biol* (2018). 14(7):e1006343–21. doi:10.1371/journal.pcbi.1006343
27. Pedro AM, Rosas F, Carhart-Harris RL, Seth A, and Adam B. Beyond Integrated Information: A Taxonomy of Information Dynamics Phenomena. (2019). arXiv preprint arXiv:1909.02297.
28. Pedro AM, Seth A, and Adam B. Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy* (2019). 21(1):17. doi:10.3390/e21010017
29. Mueller MP. Could the Physical World Be Emergent Instead of Fundamental, and Why Should We Ask?(short Version). (2017). arXiv preprint arXiv:1712.01816.
30. Northoff G, Tsuchiya N, and Saigo H. Mathematics and the Brain. A Category Theoretic Approach to Go beyond the Neural Correlates of Consciousness. (2019). bioRxiv.
31. Oizumi M, Albantakis L, and Tononi G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* (2014). 10(5):e1003588. doi:10.1371/journal.pcbi.1003588
32. Encyclopedia of Mathematics. *Signed Measure*. Berlin: Springer (2013).
33. Anthony P. Consciousness as Integrated Information a Provisional Philosophical Critique. *J Conscious Stud* (2013). 20(1–2):180–206. doi:10.2307/25470707
34. Pedro R. *Proceedings of the Workshop on Combining Viewpoints in Quantum Theory, 19–22*. Edinburgh, UK: ICMS (2018).
35. Walter R. *Real and Complex Analysis*. London: Tata McGraw-hill education (2006).
36. Salamon D. *Measure and Integration*. London: European Mathematical Society (2016). doi:10.4171/159
37. Seth A, Adam B, and Barnett L. Causal Density and Integrated Information as Measures of Conscious Level. *Philos Trans A Math Phys Eng Sci* (1952). 369: 3748–67. doi:10.1098/rsta.2011.0079
38. Sharma G, Wu W, and Edul N. Dalal. The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations. London: COLOR Research and Application (2004).
39. Miguel Signorelli C, Wang Q, and Khan I. A Compositional Model of Consciousness Based on Consciousness-Only. (2020). arXiv preprint arXiv:2007.16138.
40. Tononi G, Boly M, Massimini M, and Koch C. Integrated Information Theory: from Consciousness to its Physical Substrate. *Nat Rev Neurosci* (2016). 17(7): 450, 61. doi:10.1038/nrn.2016.44
41. Tegmark M. Consciousness as a State of Matter. *Chaos, Solitons Fractals* (2015). 76:238–70. doi:10.1016/j.chaos.2015.03.014
42. Tegmark M. Improved Measures of Integrated Information. *PLoS Comput Biol* (2016). 12(11). doi:10.1371/journal.pcbi.1005123
43. Tsuchiya N, Haun A, Cohen D, and Oizumi M. *Empirical Tests of the Integrated Information Theory of consciousnessThe Return of Consciousness: A New Science on Old Questions*. London: Axel and Margaret Ax. son Johnson Foundation (2016). p. 349–74.
44. Tull S, and Kleiner J. Integrated Information in Process Theories. *J Cognit Sci* (2021). 22:135–55.
45. Tononi G. An Information Integration Theory of Consciousness. *BMC Neurosci* (2004). 5(1):42. doi:10.1186/1471-2202-5-42
46. Tononi G. Consciousness, Information Integration, and the Brain. *Prog Brain Res* (2005). 150:109–26. doi:10.1016/s0079-6123(05)50009-8
47. Tononi G. Consciousness as Integrated Information: a Provisional Manifesto. *Biol Bull* (2008). 215(3):216–42. doi:10.2307/25470707
48. Tononi G. Integrated Information Theory. *Scholarpedia* (2015). 10(1):4164. doi:10.4249/scholarpedia.4164
49. Tsuchiya N, Taguchi S, and Saigo H. Using Category Theory to Assess the Relationship between Consciousness and Integrated Information Theory. *Neurosci Res* (2016). 107(1–7):133. doi:10.1016/j.neures.2015.12.007
50. Zanardi P, Tomka M, and Venuti LC. Quantum Integrated Information Theory. (2018). arXiv preprint arXiv:1806.01421, 2018 Comparison with Standard Presentation of IIT 3.0.

Conflict of Interest: Author ST was employed by company Cambridge Quantum Computing Limited.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A: COMPARISON WITH STANDARD PRESENTATION OF IIT 3.0

In **Section 9**, we have defined the system class and cause-effect repertoires which underlie classical IIT. The goal of this appendix is to explain in detail why applying our definition of the IIT algorithm yields IIT 3.0 defined by Tononi et al. In doing so, we will mainly refer to the terminology used in [25, 26, 31, 48]. We remark that a particularly detailed presentation of the algorithm of the theory, and of how the cause and effect repertoire are calculated, is given in the Supplementary Material S1 of [26].

A.1 Physical Systems

The systems of classical IIT are given in **Section 9.1**. They are often represented as graphs whose nodes are the elements $S_1, \ldots, S_n$ and edges represent functional dependence, thus describing the time evolution of the system as a whole, which we have taken as primitive in **Eq. 18**. This is similar to the presentation of the theory in terms of a transition probability function

$$p : \mathrm{St}(S) \times \mathrm{St}(S) \to [0, 1]$$

in [25]. For each probability distribution $\tilde{p}$ over $\mathrm{St}(S)$, this relates to our time evolution operator T via

$$T(\tilde{p})[v] := \sum_{w \in \mathrm{St}(S)} p(v, w)\, \tilde{p}(w).$$

A.2 Cause-Effect Repertoires

In contemporary presentations of the theory ([25], p. 14] or [48]), the effect repertoire is defined as

$$p_{\text{effect}}(z_i, m_t) := \frac{1}{|\Omega_{M^c}|} \sum_{m^c \in \Omega_{M^c}} p(z_i | do(m_t, m^c)) \quad z_i \in \Omega_{Z_i} \tag{35}$$

and

$$p_{\text{effect}}(z, m_t) := \prod_{i=1}^{|z|} p_{\text{effect}}(z_i, m_t). \tag{36}$$

Here, m_t denotes a state of the mechanism M at time t. M^c denotes the complement of the mechanism, denoted in our case as $M^{\perp}$, Ω_{M^c} denotes the state space of the complement, and m^c an element thereof. Z_i denotes an element of the purview Z (designated by P in our case), Ω_{Z_i} denotes the state space of this element, z_i a state of this element and z a state of the whole purview. $|\Omega_{M^c}|$ denotes the cardinality of the state space of M^c, and $|z|$ equals the number of elements in the purview. Finally, the expression $\mathrm{do}(m_t, m^c)$ denotes a variant of the so-called "do-operator". It indicates that the state of the system, here at time t, is to be set to the term in brackets. This is called *perturbing the system* into the state (m_t, m^c). The notation $p(z_i | \mathrm{do}(m_t, m^c))$ then gives the probability of finding the purview element in the state z_i at time $t + 1$ given that the system is prepared in the state (m_t, m^c) at time t.

In our notation, the right hand side of **Eq. 35** is exactly given by the right-hand side of **Eq. 25**, i.e. $\mathrm{eff}'_s(M, P_i)$. The system is prepared in a uniform distribution on M^c (described by the sum and prefactor in **Eq. 35**) and with the restriction s_M of the system state, here denoted by m_t, on M. Subsequently, T is applied to evolve the system to time $t + 1$, and the marginalization $\langle P_i^{\perp} |$ throws away all parts of the states except those of the purview element P_i (denoted above as Z_i). In total, **Eq. 25** is a probability distribution on the states of the purview element. When evaluating this probability distribution at one particular state z_i of the element, one obtains the same numerical value as **Eq. 35**. Finally, taking the product in **Eq. 36** corresponds exactly to taking the product in **Eq. 26**.

Similarly, the cause repertoire is defined as ([25], p. 14] or [48])

$$p_{\text{cause}}(z | m_{i,t}) := \frac{\sum_{z^c \in \Omega_{Z^c}} p(m_{i,t} | \mathrm{do}(z, z^c))}{\sum_{s \in \Omega_S} p(m_{i,t} | \mathrm{do}(s))} \quad z \in \Omega_{Z_{t-1}} \tag{37}$$

and

$$p_{\text{cause}}(z | m_t) := \frac{1}{K} \prod_{i=1}^{|m_t|} p_{\text{cause}}(z | m_{i,t}), \tag{38}$$

where m_i denotes the state of one element of the mechanism M, with the subscript t indicating that the state is considered at time t. Z again denotes a purview, z is a state of the purview and $\Omega_{Z_{t-1}}$ denotes the state space of the purview, where the subscript indicates that the state is considered at time $t - 1$. K denotes a normalization constant and $|m_t|$ gives the number of elements in M.

Here, the whole right hand side of **Eq. 37** gives the probability of finding the purview in state z at time $t - 1$ if the system is prepared in state $m_{i,t}$ at time t. In our terminology this same distribution is given by **Eq. 27**, where λ is the denominator in **Eq. 37**. Taking the product of these distributions and re-normalising is then precisely **Eq. 28**.

As a result, the cause and effect repertoire in the sense of [31] correspond precisely in our notation to $\mathrm{caus}'_s(M, P)$ and $\mathrm{eff}'_s(M, P)$, each being distributions over $\mathrm{St}(P)$. In (Supplementary Material S1 of [26]), it is explained that these need to be extended by the unconstrained repertoires before being used in the IIT algorithm, which in our formalization is done in **Eq. 29**, so that the cause-effect repertoires are now distributions over $\mathrm{St}(S)$. These are in fact precisely what are called the *extended* cause and effect repertoires or *expansion to full state space* of the repertoires in [31].

The behavior of the cause- and effect-repertoires when decomposing a system is described, in our formalism, by decompositions (Definition 5). Hence a decomposition $z \in \mathbb{D}_S$ is what is called a *parition* in the classical formalism. For the case of classical IIT, a decomposition is given precisely by a partition of the set of elements of a system, and the cause-effect repertoires belonging to the decomposition are defined in **Eq. 30**, which corresponds exactly to the definition

$$p^{\text{cut}}_{\text{cause}}(z | m_t) = p_{\text{cause}}(z^{(1)} | m_t^{(1)}) \times p_{\text{cause}}(z^{(2)} | m_t^{(2)})$$

in [25], when expanded to the full state space, and equally so for the effect repertoire.

A.3 Algorithm: Mechanism Level

Next, we explicitly unpack our form of the IIT algorithm to see how it compares in the case of classical IIT with [31]. In our formalism, the integrated information φ of a mechanism M of system S when in state s is

$$\varphi^{\max}(M) = \|\mathbb{C}_{S,s}(M)\| \tag{39}$$

defined in **Eq. 10**. This definition conjoins several steps in the definition of classical IIT. To explain why it corresponds exactly to classical IIT, we disentangle this definition step by step.

First, consider $\text{caus}_s(M,P)$ in **Eq. 9**. This is, by definition, a decomposition map. The calculation of the integration level of this decomposition map, cf. **Eq. 5**, amounts to comparing $\text{caus}_s(M,P)$ to the cause-effect repertoire associated with every decomposition using the metric of the target space $\mathbb{PE}(S)$, which for classical IIT is defined in **Eq. 24** and Example 3, so that the metric d used for comparison is indeed the Earth Mover's Distance. Since cause-effect repertoires have, by definition, unit intensity, the factor r in the definition (1) of the metric does not play a role at this stage. Therefore, the integration level of $\text{caus}_s(M,P)$ is exactly the *integrated cause information*, denoted as

$$\varphi^{\text{MIP}}_{\text{cause}}(y_t, Z_{t-1})$$

in [48], where y_t denotes the (induced state of the) mechanism M in this notation, and Z_{t-1} denotes the purview P. Similarly, the integration level of $\text{eff}_s(M,P)$ is exactly the *integrated effect information*, denoted as

$$\varphi^{\text{MIP}}_{\text{effect}}(y_t, Z_{t+1}).$$

The integration scaling in **Eq. 10** simply changes the intensity of an element of $\mathbb{PE}(S)$ to match the integration level, using the scalar multiplication, which is important for the system level definitions. When applied to $\text{caus}_s(M,P)$, this would result in an element of $\mathbb{PE}(S)$ whose intensity is precisely $\varphi^{\text{MIP}}_{\text{cause}}(y_t, Z_{t-1})$.

Consider now the collections (9) of decomposition maps. Applying Definition 9, the core of $\text{caus}_s(M)$ is that purview P which gives the decomposition $\text{caus}_s(M,P)$ with the highest integration level, i.e. with the highest $\varphi^{\text{MIP}}_{\text{cause}}(y_t, Z_{t-1})$. This is called the *core cause* P^c of M, and similarly the core of $\text{eff}_s(M)$ is called the *core effect* P^e of M.

Finally, to fully account for **Eq. 10**, we note that the integration scaling of a pair of decomposition maps rescales both elements to the minimum of the two integration levels. Hence the integration scaling of the pair $(\text{caus}_s(M,P), \text{eff}(M,P'))$ fixes the scalar value of both elements to be exactly the *integrated information*, denoted as

$$\varphi(y_t, Z_{t\pm1}) = \min\left(\varphi^{\text{MIP}}_{\text{cause}}, \varphi^{\text{MIP}}_{\text{effect}}\right)$$

in [48], where $P = Z_{t+1}$ and $P' = Z_{t-1}$.

In summary, the following operations are combined in **Eq. 10**. The core of $(\text{caus}_s(M), \text{eff}_s(M))$ picks out the core cause P^c and core effect P^e. The core integration scaling subsequently considers the pair $(\text{caus}_s(M,P^c), \text{eff}(M,P^e))$, called *maximally irreducible cause-effect repertoire*, and determines the integration level of each by analysing the behavior with respect to decompositions. Finally, it rescales both to the minimum of the integration levels. Thus it gives exactly what is called $\varphi^{\max}$ in [48]. Using, finally, the definition of the intensity of the product $\mathbb{PE}(S) \times \mathbb{PE}(S)$ in Definition 4, this implies (39). The concept of M in our formalization is given by the tuple

$$\mathbb{C}_{S,s}(M) := \left(\left(\text{caus}_s(M,P^c), \varphi^{\max}(M)\right), \left(\text{eff}_s(M,P^e), \varphi^{\max}(M)\right)\right)$$

i.e., the pair of maximally irreducible repertoires scaled by $\varphi^{\max}(M)$. This is equivalent to what is called a *concept*, or sometimes *quale sensu stricto*, in classcial IIT [48], and denoted as $q(y_t)$.

We finally remark that it is also possible in classical IIT that a cause repertoire value $\text{caus}_s(M,P)$ vanishes (Remark 12). In our formalization, it would hence be represented by $(\omega_S, 0)$ in $\mathbb{PE}(S)$, so that $d(\text{caus}_s(M,P), q) = 0$ for all $q \in \mathbb{E}(S)$ according to (1), which certainly ensures that $\varphi^{\text{MIP}}_{\text{cause}}(M,P) = 0$.

A.4 Algorithm: System Level

We finally explain how the system level definitions correspond to the usual definition of classical IIT.

The Q-shape $\mathbb{Q}_s(S)$ is the collection of all concepts specified by the mechanisms of a system. Since each concept has intensity given by the corresponding integrated information of the mechanism, this makes $\mathbb{Q}_s(S)$ what is usually called the *conceptual structure* or *cause-effect structure*. In [31], one does not include a concept for any mechanism M with $\varphi^{\max}(M) = 0$. This manual exclusion is unnecessary in our case because the mathematical structure of experience spaces implies that mechanisms with $\varphi^{\max}(M) = 0$ should be interpreted as having no conscious experience, and the algorithm in fact implies that they have 'no effect'. Indeed we will now see that they do not contribute to the distances in $\mathbb{E}(S)$ or any Φ values, and so we do not manually exclude them.

When comparing $\mathbb{Q}_s(S)$ with the Q-shape **Eq. 13** obtained after replacing S by any of its cuts, it is important to note that both are elements of $\mathbb{E}(S)$ defined in **Eq. 12**, which is a product of experience spaces. According to Definition 4, the distance function on this product is

$$d(\mathbb{Q}_s(S), \mathbb{Q}_s(S^z)) := \sum_{M \in \text{Sub}(S)} d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M)).$$

Using Definition 3 and the fact that each concept's intensity is $\varphi^{\max}(M)$ according to the mechanism level definitions, each distance $d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M))$ is equal to

$$\begin{aligned}&\varphi^{\max}(M) \cdot \{d[\text{caus}_s(M, P^c_M), \text{caus}^z_s(M, P^{z,c}_M)]\\ &+ d[\text{eff}_s(M, P^e_M), \text{eff}^z_s(M, P^{z,e}_M)]\},\end{aligned} \tag{40}$$

where $\varphi^{\max}(M)$ denotes the integrated information of the concept in the original system S, and where the right-hand cause and effect repertoires are those of S^z at its own core causes and effects for M. The factor $\varphi^{\max}(M)$ ensures that the distance used here corresponds precisely to the distance used in [31], there called the *extended Earth Mover's Distance*. If the integrated information $\varphi^{\max}(M)$ of a mechanism is non-zero, it follows that

$d(\mathbb{C}_{S,s}(M), \mathbb{C}_{S^z,s^z}(M)) = 0$ as mentioned above, so that this concept does not contribute.

We remark that in Supplementary Material S1 of [26], an additional step is mentioned which is not described in any of the other papers we consider. Namely, if the integrated information of a mechanism is non-zero before cutting but zero after cutting, what is compared is not the distance of the corresponding concepts as in **Eq. 40**, but in fact the distance of the original concept with a special null concept, defined to be the unconstrained repertoire of the cut system. We have not included this step in our definitions, but it could be included by adding a choice of distinguished point to Example 3 and redefining the metric correspondingly.

In **Eq. 14** the above comparison is being conducted for every subsystem of a system S. The subsystems of S are what is called *candidate systems* in [31], and which describe that 'part' of the system that is going to be conscious according to the theory (cf. below). Crucially, candidate systems are subsystems of S, whose time evolution is defined in **Eq. 22**. This definition ensures that the state of the elements of S which are not part of the candidate system are fixed in their current state, i.e., constitute *background conditions* as required in the contemporary version of classcial IIT [26].

Eq. 14 then compares the Q-shape of every candidate system to the Q-shape of all of its cuts, using the distance function described above, where the cuts are defined in **Eq. 23**. The cut system with the smallest distance gives the system-level *minimum information partition* and the *integrated (conceptual) information* of that candidate system, denoted as $\Phi(x_t)$ in [48].

The core integration scaling finally picks out that candidate system with the largest integrated information value. This candidate system is the *major complex* M of S, the part of S which is conscious according to the theory as part of the *exclusion postulate* of IIT. Its Q-shape is the *maximally irreducible conceptual structure (MICS)*, also called *quale sensu lato*. The overall *integrated conceptual information* is, finally, simply the intensity of $\mathbb{E}(S,s)$ as defined in **Eq. 14**,

$$\Phi(S,s) = \mathbb{E}(S,s).$$

A.5 Constellation in Qualia Space

Expanding our definitions, and denoting the major complex by M with state $m = s|_M$, in our terminology the experience of system S state s is

$$\mathbb{E}(S,s) := \frac{\Phi(M,m)}{||\mathbb{Q}_m(M)||} \cdot \mathbb{Q}_m(M). \tag{41}$$

This encodes the Q-shape $\mathbb{Q}_m(M)$, i.e. the maximally irreducible conceptual structure of the major complex, sometimes called *quale sensu lato*, which is taken to describe the quality of conscious experience. By construction it also encodes the integrated conceptual information of the major complex, which captures its intensity, since we have $||\mathbb{E}(S,s)|| = \Phi(M,m)$. The rescaling of $\mathbb{Q}_m(M)$ in **Eq. 41** leaves the relative intensities of the concepts in the MICS intact. Thus $\mathbb{E}(S,s)$ is the *constellation of concepts in qualia space* $\mathbb{E}(M)$ of [31].

Integrated Information in Process Theories: Towards Categorical IIT

Sean Tull[1] and Johannes Kleiner[2]

[1]*Cambridge Quantum Computing*
[2]*Munich Center for Mathematical Philosophy*
sean.tull@cambridgequantum.com, johannes.kleiner@lmu.de

Abstract

We demonstrate how integrated information and other key notions from Tononi et al.'s Integrated Information Theory (IIT) can be studied within the simple graphical language of process theories (symmetric monoidal categories). This allows IIT to be generalised to a broad range of physical theories, including as a special case the Quantum IIT of Zanardi, Tomka and Venuti, and sets the foundation for a categorical definition of IIT.

Keywords: *Integrated Information Theory, Process theory, Monoidal Category, Consciousness, Quantum Integrated Information Theory*

1. Introduction

Integrated Information Theory (IIT) is a theory of consciousness proposed and developed by Giulio Tononi and collaborators (Tononi, 2008; Oizumi et al., 2014). Originally defined in terms of a numerical measure Φ representing the level of phenomenal consciousness of a system (Tononi, 2004; Mediano et al., 2019), the most recent version of the theory, IIT 3.0, now employs an algorithm which claims to determine in addition which part of a system is conscious, and what it is conscious of.

Received 28th February 2021; Revised 25th May 2021; Accepted 30th May 2021
Journal of Cognitive Science 22(2): 92-123 June 2021

In this article we show how the key concepts of IIT, including systems, integration and causation, can be studied naturally in the language of physical *process theories*, which are mathematically described as *symmetric monoidal categories*. Process theories come with an intuitive but rigorous graphical calculus (Selinger, 2011) which allows us to present many aspects of IIT in a simple pictorial fashion.

The constructions we provide in this article can be applied to any suitable process theory to yield a notion of *generalised IIT* as defined by the authors in a companion article (Kleiner and Tull, 2021). This allows us to extend IIT to new physical settings. As special cases, choosing the process theory of classical probabilistic processes essentially yields the usual IIT 3.0 in the sense of (Oizumi et al., 2014). Starting instead from the theory of quantum processes gives the *Quantum Integrated Information Theory* defined by Zanardi, Tomka and Venuti (Zanardi et al., 2018), which was another motivation for this work.

Independently of consciousness itself, our constructions provide a possible foundation for a general theory of integrated or 'holistic' behaviour within process theories, i.e. monoidal categories, which may be of interest to a broad range of fields. For example, neural net-like systems that achieve a task using a high degree of integration should be more efficient than fully modular ones, in that they require fewer neurons for the same task, and indeed integrated behaviour has been shown to evolve in simple models of biological organisms (Albantakis et al., 2014). The methods of IIT have been applied generally in the study of integration in information processing systems, including treatments of autonomy (Marshall et al., 2017), causation (Albantakis et al., 2017), and state differentiation (Marshall et al., 2016).

1.1 Background: Mathematical Consciousness Science

The background for our work is in the growing field of Mathematical Consciousness Science (MCS), which aims to apply formal and mathematical tools in order to resolve open problems in the scientific study of consciousness. One major goal thereby is to expose and improve the mathematical structure of neuroscientific theories of consciousness so as to allow quantifiable comparison between competing models, generate novel experimental predictions, and to provide a thorough foundation for further development

and combination of theories. More foundationally, it aims to uncover how consciousness relates to the physical world in terms of empirically grounded and philosophically motivated scientific theories. Progress in this direction is essential for resolving medical challenges (most notably, improving the understanding of neurological, psychiatric and psychological disorders (Michel et al., 2019)) and ethical reasons (for example the detection of consciousness in anesthetized or non-communicating patients (Alkire et al., 2008; Fink et al., 2018)), and could generate new advances in AI (artificial implementation of consciousness-related functions, for example (McDermott, 2007)).

A crucial cornerstone in this program is the representation of conscious experience in terms of a mathematical spaces, and to expound theories of consciousness as mappings from a mathematical description of physical systems to these spaces. Early precursors of the former are quality spaces (Beals et al., 1968; Clark, 1996, 2000) which make use of just noticeable difference between stimuli to construct a representation of mental qualities and similarities between them. In the companion article (Kleiner and Tull, 2021), we provide a definition of an *experience space* that builds upon quality spaces while being geared at precisely what is required to flash out IIT as a mathematical mapping of the just-mentioned kind.

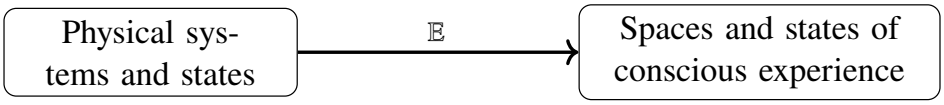
This contributes to the exploration and application of category theory as a framework for theories of consciousness (Tsuchiya et al., 2016; Northoff et al., 2019; Ehresmann, 2012). Category theory itself provides a natural language for describing mappings between scientific domains, such as domains of physical systems and those modelling phenomenal experiences. Its emphasis on processes between systems in particular makes it ideal for describing theories and experimental findings which relate consciousness to dynamical processes, as discussed for example in (Fekete and Edelman, 2011; Wiese and Friston, 2020; Grindrod, 2018). The use of monoidal categories in this article additionally allows us to treat compositional aspects of systems and processes, which are central to theories such as IIT.

1.2 A Primer on Integrated Information Theory

Though the majority of the article is self-contained and requires no prior knowledge of the theory, for context we include here a short introduction

to IIT 3.0 (Oizumi et al., 2014), as formulated in its general form in our companion article (Kleiner and Tull, 2021) to which we refer for a more detailed presentation of the theory.

Any generalised IIT, including IIT 3.0, takes as input a given class of physical systems S, each with a given state space $\mathrm{St}(S)$, and specifies a map $\mathbb{E}$ which provides each system with a space describing its possible conscious experiences. Additionally, for each state $s \in \mathrm{St}(S)$ the theory specifies a particular experience $\mathbb{E}(s) \in \mathbb{E}(S)$ which the system will have in that state:



In IIT 3.0 the nature of this mapping derives from a number of essential properties–so called 'axioms'–which are postulated to characterize every conscious experience. Next to integration and information, these axioms include intrinsic existence, composition and exclusion (Tononi, 2015). These axioms are being translated into formal requirements. To this end, comparably simple physical systems are considered. These consist of a set of elements (or 'nodes'), each usually with only two states (on or off), and come with a discrete Markovian time evolution which is often described via a given causal graph. The prototypical example would be a human brain, in which the nodes represent neurons and their firing. The result of the translation process is the algorithm of IIT 3.0, i.e. the map $\mathbb{E}$ when applied to classical physical systems.

Starting from such a system S along with its current state s, the theory then specifies a set of probability distributions known as the *cause-effect repertoire*. For each pair of subsystems M, P ('mechanism' and 'purview') of S, the cause repertoire $\mathsf{caus}(M, P)$ is a distribution specifying how the current state of M constrains the state of P in the previous time-step, and similarly the effect repertoire $\mathsf{eff}(M, P)$ addresses the next time-step instead.

In the IIT algorithm one goes on how to calculate how 'integrated' each of these repertoires are by comparing them against repertoires obtained instead by 'cutting' the (evolution of the) system into various parts, by removing causal connections between them. For each mechanism M one determines which purviews give the most integrated values of $\mathsf{caus}(M, P)$ and $\mathsf{eff}(M, P)$, and these repertoire values (along with their level of integration) determine a *concept* for that mechanism. The weighted collection of these concepts

determines the entity $\mathbb{E}(s)$, also known as the *Q-shape* of the system, which is claimed to specify its total conscious experience. In particular this Q-Shape comes with its own level of integration, denoted $\Phi(s)$, which describes 'how conscious' the system is as a whole. A final 'exclusion' step enforces that only the subsystem of S with the highest Φ value will in fact be conscious.

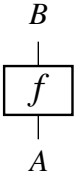
In the article (Kleiner and Tull, 2021) we show how to define a broad class of generalisations of IIT, in which for example the repertoires need no longer be described by probability distributions, but the states of a general physical theory. In the present article we describe how such IITs may be defined starting from any physical process theory. To do so we define the key notions of any IIT within such a setting, namely causal relations and their integration.

1.3 Structure of Article

The article is structured as follows. We introduce process theories in Section 2 and then use them to describe the key notions from IIT – decompositions of objects (Section 3), systems (Section 4) and cause and effect repertoires (Section 5). We summarise how to define a generalised IIT from a process theory in Section 6 before giving examples in Section 7 and discussing future work in Section 8. The appendix contains some initial steps in developing a general study of integration in monoidal categories.

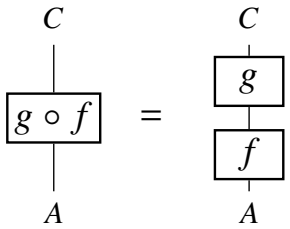
2. Process Theories

We begin by introducing the framework of *process theories* used throughout this work; for more detailed introductions we refer to (Coecke and Paquette, 2010; Coecke and Kissinger, 2017). The basic ingredients of such a theory are *objects* and *processes* between them. We depict a process from the object A to the object B as a box:

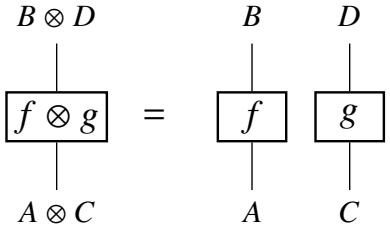


These processes may be *composed* together to form new ones in several ways. Firstly, given a process such as f above, and any other process g from

B to C, we may compose them 'in sequence' to form a new one from A to C, denoted:

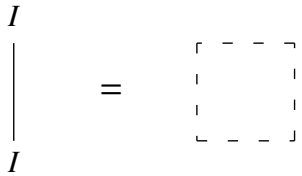


Secondly, we may compose processes in parallel. Any two objects A, B may be combined into a single object $A \otimes B$. Moreover any processes f from A to B, and g from C to D may be placed 'side-by-side' to form a new process:



from $A \otimes C$ to $B \otimes D$. More generally, by combining these operations, many processes may all be plugged together to form more complex diagrams describing a single composite process.

As a convenience, any process theory is taken to come with the following. Firstly, any object A come with an *identity process*, depicted as a blank wire on A, which 'does nothing' in that composing with it via $\circ$ leaves any process as it is. Secondly, it has a *trivial object*, denoted I, which leaves objects alone when combining under $\otimes$. We depict I as empty space:

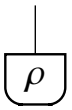


Finally, we formally assume the presence of a special process $\times$ which allows us to 'swap' any pair of wires over each other, along with a set of rules saying roughly that diagrams in the above sense are well-defined.

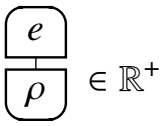
Mathematically, all of this is summarised by saying that a process theory is precisely a *symmetric monoidal category* $(\mathbf{C}, \otimes, I)$ with the processes as

its *morphisms*. Our diagrammatic rules correspond to the precise *graphical calculus* for reasoning in such categories (Selinger, 2011).

We will often wish to refer to some special kinds of processes. Processes with 'no input' in diagrams (and so formally with input object I) are called *states*, and can be thought of as 'preparations' of the physical system given by their output object:



Processes with no output, called *effects*, may be thought of as 'observations' we may record on our system. Finally, processes with neither input nor output are called *scalars*. It is common for theories to come with a *probabilistic* interpretation meaning that each of their scalars p correspond to a probability, or more generally an 'unnormalised probability' $p \in \mathbb{R}^+$, with $r \otimes s = r \cdot s$ for scalars and the empty diagram given by 1. In particular, the composition of a state with an effect



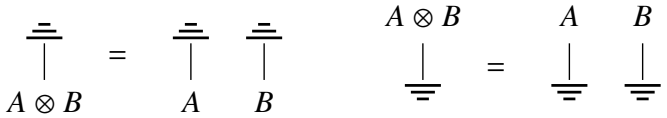
$\in \mathbb{R}^+$

corresponds to the 'probability' of observing the effect e in the state ρ. Such 'generalised probabilistic theories' are a major focus of study in the foundations of physics (Barrett, 2007).

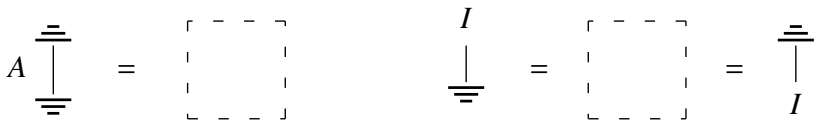
The theories we consider here will often come with further structure giving them a physical interpretation. Firstly, every object will come with a distinguished *discarding* effect depicted

which we think of as the process of simply 'throwing away' or 'ignoring' a physical system. Similarly, every object should come with a distinguished *completely mixed state* depicted as

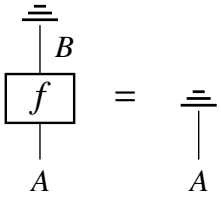
which corresponds to preparing the object in a maximally 'noisy' or 'random' state. These processes should satisfy



as well as



for all objects A, B. We then define a process f to be *causal* when it satisfies

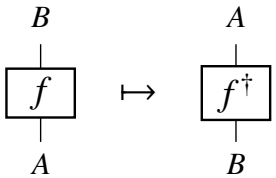


or similarly as *co-causal* if it preserves $\perp\!\!\!\perp$. Discarding processes are in fact closely related to physical notions of causality; see for example (Coecke, 2014; Chiribella et al., 2010).

In such a probabilistic theory there is a unique process between any two objects, the *zero process* 0, such that composing any process via $\circ, \otimes$ with 0 always yields 0.

At times we will assume our process theory also comes with a way of describing how similar any two causal states are. This amounts to a choice of *distance function* on the set $\mathrm{St}_{\mathsf{c}}(A)$ of causal states of each object A, providing a value $d(a, b) \in \mathbb{R}^+$ for each $a, b \in \mathrm{St}_{\mathsf{c}}(A)$. Often this map d will satisfy the axioms of a metric, but this is not required.

Our main examples of process theories will come with a notable extra feature, though this will not be necessary for our approach. In many theories it is possible to 'reverse' any process, in that for any process f there is another $f^\dagger$ in the opposite direction. We say a process theory has a *dagger* when it comes with such a mapping



which preserves composition and identity maps in an appropriate sense, and satisfies $f^{\dagger\dagger} = f$ for all f. The presence of a dagger is a common starting point in categorical approaches to quantum theory; see e.g. (Abramsky and Coecke, 2004; Selinger, 2007).

Let us now meet our main examples of process theories with the above features.

Example 1 (**Classical Probabilistic Processes**) In the process theory **Class** of finite-dimensional probabilistic classical physics, the objects are finite sets $A, B, C, \ldots$ and the processes f from A to B are functions sending each element $a \in A$ to a 'unnormalised probability distribution' over the elements of B, i.e functions $f\colon A \times B \to \mathbb{R}^+$. Composition of f from A to B and g from B to C is defined by

$$(g \circ f)(a, c) = \sum_{b \in B} f(a, b) \cdot g(b, c)$$

In this theory the trivial object is the singleton set $I = \{\star\}$, with $\otimes$ given by the Cartesian product $A \times B$ and $(f \times g)(a, c)(b, d) = f(a, b) \cdot g(c, d)$. This theory is probabilistic, with scalars $r \in \mathbb{R}^+$.

Here $\overset{\cdot}{\top}_A$ is the unique effect with $\overset{\cdot}{\top}_A(a) = 1$ for all $a \in A$. A process f is causal whenever it is stochastic, i.e. sends each element $a \in A$ to a (normalised) probability distribution over the elements of B. Applying the process $\overset{\cdot}{\top}$ to some output wire of a process corresponds to *marginalising* over the set which is discarded.

States of an object are '$\mathbb{R}^+$-distributions' over their elements, while causal states are normalised ones, i.e. probability distributions. The completely mixed state $\overset{\cdot}{\perp}_A$ is the uniform probability distribution, with $\overset{\cdot}{\perp}_A(a) = \frac{1}{|A|}$ for all $a \in A$. This theory also has a dagger by $f^\dagger(b, a) = f(a, b)$.

Similarly we define another process theory $\mathbf{Class}_\mathsf{m}$, in the same way, but with objects now being finite *metric spaces* (A, d). Each object A now comes with a metric d on its underlying set, with $A \otimes B = A \times B$ having the product metric. For each object A we extend d to a metric d_W on probability distributions over A, i.e. causal states of A, called the *Wasserstein metric* or *Earth Mover's Distance* (EMD), definable e.g. by

$$d_W(s, t) := \sup_f \{\sum_{a \in A} f(a) \cdot s(a) - \sum_{a \in A} f(a) \cdot t(a)\}$$

where the suprema is taken over all functions f satisfying $|f(a) - f(b)| \leq d(a, b)$ for all a, b. **Class** itself may be given a metric on causal states in the same way by taking each object A to have metric $d(a, b) = 1 - \delta_{a,b}$.

Example 2 (**Quantum Processes**) In the process theory **Quant** the objects are finite-dimensional complex Hilbert spaces $\mathcal{H}, \mathcal{K}, \ldots$ and the processes from $\mathcal{H}$ to $\mathcal{K}$ are *completely positive maps* $f\colon B(\mathcal{H}) \to B(\mathcal{K})$ between their spaces of operators. Here $I = \mathbb{C}$ and $\otimes$ is the usual tensor product of Hilbert spaces and maps. States ρ of an object $\mathcal{H}$ may be identified with (unnormalised) *density matrices*, i.e. quantum states in the usual sense, as may effects. The effect $\overline{\top}$ sends each operator $a \in B(\mathcal{H})$ to its *trace* $\mathrm{Tr}(a)$, and $\overline{\bot}$ is the maximally mixed state on $\mathcal{H}$, with density matrix $\frac{1}{\dim(\mathcal{H})} 1_{\mathcal{H}}$. Here a process is causal precisely when it is trace-preserving, and the dagger is given by the Hermitian adjoint.

Example 3 (**Quantum-Classical Processes**) To combine **Class** and **Quant** we may use the theory **CStar** whose objects are finite-dimensional C^*-*algebras* $A, B, \ldots$ and processes are completely positive maps $f\colon A \to B$, with $\otimes$ given by the standard tensor product, $I = \mathbb{C}$ and the dagger again by the Hermitian adjoint. Here $\overline{\top}$ sends each element $a \in A$ to its trace $\mathrm{Tr}(a) \in \mathbb{C}$, while $\overline{\bot}$ corresponds to the rescaling $\frac{1}{d}1$ of the element $1 \in A$, where $\mathrm{Tr}(1) = d$. Each C^*-algebra comes with a metric induced by its norm, providing a metric on states in the theory.

Class may be identified with the sub-theory of **CStar** containing the commutative algebras, and **Quant** with those of the form $B(\mathcal{H})$ for some Hilbert space $\mathcal{H}$. More general algebras are 'quantum-classical', being given by direct sums of quantum algebras.

3. Decompositions

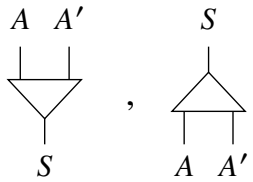
A central aspect of IIT is evaluating the level of integration of a process, and particularly of a state of some object. To do so we must compare the object in question against ways it may be *decomposed*, as follows.

Firstly, recall that a process f from A to B is an *isomorphism* when there is some (unique) f^{-1} from B to A for which $f^{-1} \circ f$ and $f \circ f^{-1}$ are both identities. We write $A \simeq B$ when such an isomorphism exists.

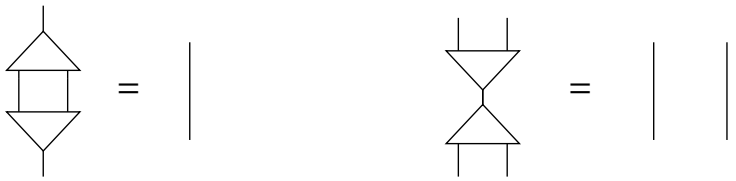
Definition 4 *In any process theory, a* decomposition *of an object S is a pair of objects A, A' along with an isomorphism $S \simeq A \otimes A'$.*

In a process theory with $\fatslash, \fatbslash$ *we will always consider decompositions whose isomorphisms are causal and co-causal. We also assume that decomposition isomorphisms preserve any distances between causal states.*

For short we often denote such a decomposition simply by (A, A') and depict its isomorphism and inverse by

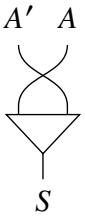


respectively. The fact that they form an isomorphism means that



One can go on to develop a general study of decompositions in process theories. Here we just note some of the basics, for more see Appendix A.

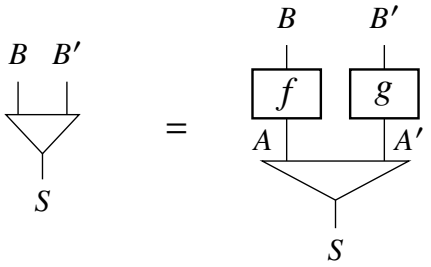
Firstly, any decomposition has an induced *complement* decomposition $(A, A')^{\perp} := (A', A)$, with isomorphism given by swapping its components:



All decompositions then satisfy $(A, A')^{\perp\perp} = (A, A')$. Moreover, any object always S always comes with *trivial decompositions* denoted $1 := (S, I)$ and $0 := (I, S)$ with $0 = 1^{\perp}$. Drawing either of their isomorphisms would just mean drawing a blank wire labelled by S.

It is also useful to note when two decompositions of an object are 'essentially the same'. We write $(A, A') \sim (B, B')$ and call both decompositions

equivalent when there exists isomorphisms f, g with



(1)

In a theory with $\dagger, \ddagger$ we require moreover that f, g are causal and co-causal.

We write $\mathbb{D}(S)$ for the set of all equivalence classes of decompositions of S under $\sim$ (we will ignore the fact that in full generality each equivalence class may be a proper class rather than a set). Often we abuse notation and denote the members of simply by (A, A') instead of as equivalence classes $[(A, A')]_\sim$. It is easy to see that if two decompositions are equivalent then so are their complements, so that $(-)^\perp$ is well-defined on $\mathbb{D}(S)$.

Definition 5 *By a* decomposition set *of an object S in a process theory we mean a subset $\mathbb{D}$ of $\mathbb{D}(S)$ containing 1 and closed under $(-)^\perp$.*

Given any decomposition set $\mathbb{D}$ of S and any $(A, A') \in \mathbb{D}$, we define the *restriction* of $\mathbb{D}$ to A via this decomposition to be the decomposition set

$$\mathbb{D}|_A := \left\{ \quad \middle| \ \exists \quad \text{s.t.} \quad \in \mathbb{D} \right\} \subseteq \mathbb{D}(A)$$

Intuitively $\mathbb{D}|_A$ consists of all decompositions of A which themselves can be extended to give a decomposition of S belonging to $\mathbb{D}$, via (A, A').

The most important examples of decomposition sets are the following.

Example 6 Let S be an object with a given isomorphism

$$S \simeq S_1 \otimes \cdots \otimes S_n$$

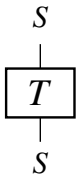
representing S as finite tensor of objects S_i which we may call *elements*. This induces a decomposition set $\mathbb{D}$ of S whose elements correspond to subsets J of the elements. For any such subset, defining $S_J := \bigotimes_J S_j$ we have a decomposition $S \simeq S_J \otimes S_{J'}$ where J' is the set of remaining elements. Then $\mathbb{D}|_{S_J}$ contains a decomposition for each $K \subseteq J$ in the same way.

Decompositions via elements as above are the only kinds appearing in classical or quantum IIT. However, more general ones allow us to treat systems which are not decomposable into any finite set of 'elementary' subsystems.

4. Systems

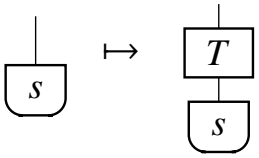
We now begin by seeing how each of the main components of IIT, or any 'generalised IIT' in the sense of (Kleiner and Tull, 2021), may be treated starting from any given process theory **C**. The focus will be on a class of *systems*, as follows.

Definition 7 *By a* system type *we mean a triple* $\underline{S} = (S, \mathbb{D}, T)$ *consisting of an object* S *with a decomposition set* $\mathbb{D}$ *and a causal process*

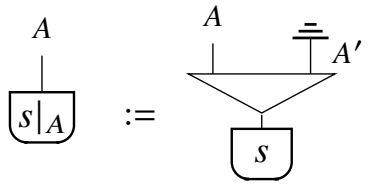


which we call its time evolution. *A* state *of* $\underline{S}$ *is simply a state of* S *in* **C**. *We typically refer to a system type simply as a* system.

The set $\mathbb{D}$ specifies the ways in which we will decompose our underlying system when assessing integration. The process T is intended to describe the way in which states of the system evolve over each single 'time-step', via



In what follows it will be useful to be able to restrict any state s of our system to the components of any decomposition $(A, A') \in \mathbb{D}$ by setting

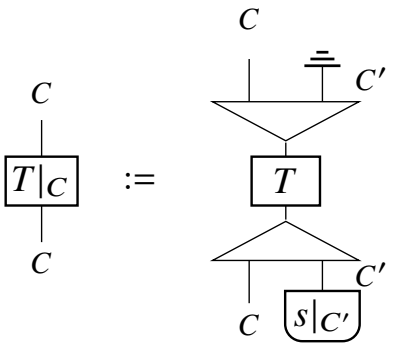


and defining $s|_{A'}$ similarly. We define the *trivial system* $\underline{I}$ to have object I, a single decomposition $1 = (I, I) = 0$, and time evolution being the identity.

4.1 Subsystems

There are several operations on systems one carries out in the context of IITs. The first is the taking of *subsystems*.

Definition 8 *For each object C belonging to some decomposition $(C, C') \in \mathbb{D}$, and each state s of $\underline{S}$, the corresponding* subsystem *of $\underline{S}$ is defined to be the system type $\underline{C}^s := (C, \mathbb{D}|_C, T|_C)$ with time evolution*



The above definition of $T|_C$ is from (Oizumi et al., 2014) and aims to capture the evolution of a state of C conditioned on the state of C' being $s|_{C'}$.

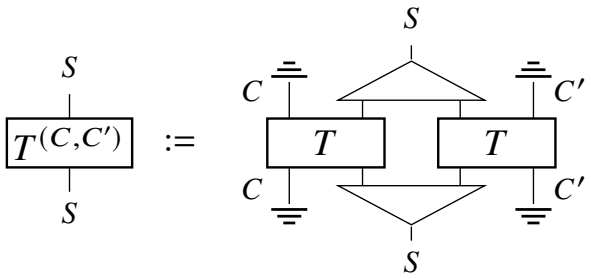
4.2 Cutting

A second important operation involves removing (some or all) causal connections between the two different components of a decomposition of a

system. For any system $\underline{S} = (S, \mathbb{D}, T)$ and decomposition $(C, C') \in \mathbb{D}$, we should be able to form a new such *cut* system of the form

$$\underline{S}^{(C,C')} = (S, \mathbb{D}, T^{(C,C')})$$

with the new evolution $T^{(C,C')}$ removing some influence between these regions. The most straightforward form of cutting is a *symmetric cut*, in which both components are fully disconnected from each other, with evolution

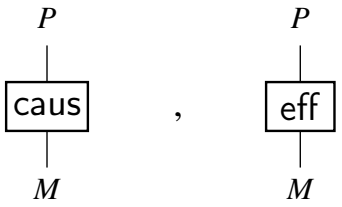


(2)

(where the triangle denotes $(C, C')^{\perp}$). However, later we will see that some IITs use additional structure to carry out alternative notions of system cut.

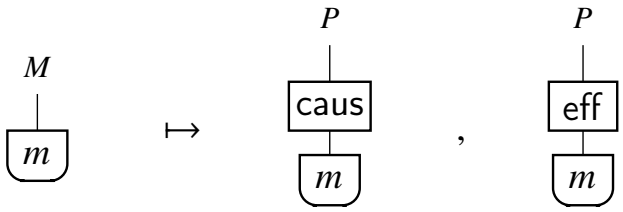
5. Cause and Effect

Central to any IIT is a notion of causal influence between any two possible subsystems of a system. These influences are captured in a pair of assignments called the *cause repertoire* and *effect repertoire* of the system. In IIT 3.0 these contain probability distributions describing how the present state of each subsystem constrains the past and future states of each other subsystem (Oizumi et al., 2014). For our purposes it suffices to note that such cause and effect repertoires amount to specifying a pair of processes



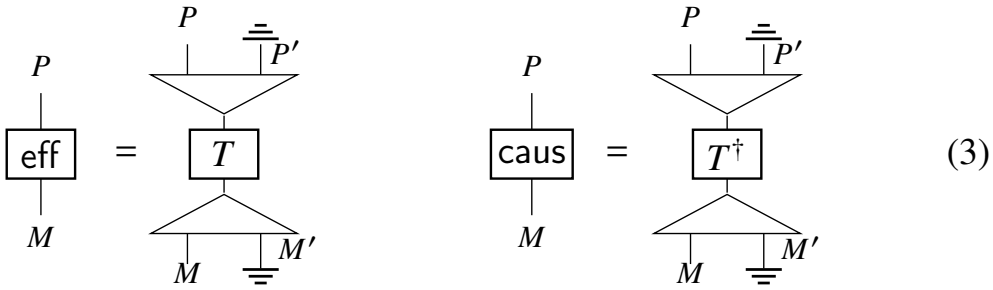
for each pair of underlying objects M, P of subsystems $\underline{M}, \underline{P}$ of $\underline{S}$ via some state s. In this setting M is typically called the 'mechanism' and P the 'purview', and the above processes should capture the way in which the

current state m of M constrains the previous or next state of P, respectively. These constraints are captured by the pair of states of P given by plugging in the 'current' state m of M:



We will additionally require the processes caus, eff to be *weakly causal* in the sense that whenever the state m is causal then each of the above states must either be causal or 0.

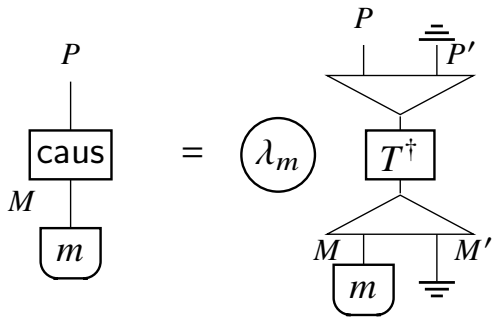
Example 9 For any process theory (resp. with a dagger) there is a simple choice of effect (resp. cause) repertoire given by



(3)

Note however that this definition of caus may not be weakly causal in our above sense if $T^\dagger$ is not causal.

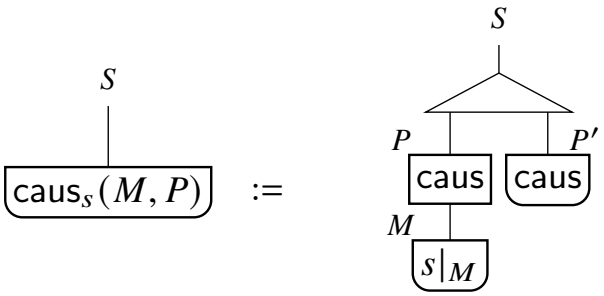
In a probabilistic process theory we should instead have that



(4)

where λ_m is the unique *normalisation* scalar for the right-hand state, making it causal if it is non-zero (and being zero otherwise). It is not in general possible to define a process caus in terms of its action on states m in this way, but this is possible for example in **Class**, **Quant** or **CStar**.

However the repertoires are specified, we will need to compare their values in a fixed state while varying P. To do so, for each state s of $\underline{S}$ and each such M, P we define the *cause repertoire at s* to be the state of S given by



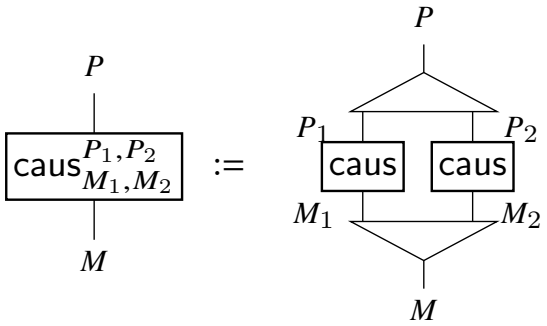
(5)

The features of this diagram have special names in (Oizumi et al., 2014); the right-hand caus state above, given by taking mechanism $M = I$, is called the *unconstrained* cause repertoire, and the whole process above $s|_M$ in the diagram is called the *extended cause repertoire* at M, P. Defining them in this way allows us to compare the repertoire values for varying M, P.

Similarly, $\mathsf{eff}_s(M,P)$, the *effect repertoire at s*, and the *unconstrained* and *extended effect repertoire* are all defined in terms of eff in the same way.

5.1 Decomposing Repertoires

In an IIT we must assess how integrated each of these repertoire values are at a given state . This involves comparing the repertoires with how they behave under decomposing each of M and P. For any decompositions $(M_1, M_2) \in \mathbb{D}|_M$ of M and $(P_1, P_2) \in \mathbb{D}|_P$ of P, the *decomposed* cause repertoire process is defined by



(6)

We then define the state $\mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M,P)$ just like (5) but replacing caus with the process (6). We decompose the effect repertoire in just the same way in terms of eff.

6. Generalised IITs

In summary, let **C** be a process theory coming with the features $\dashv\!\vdash, \perp\!\!\!\perp, d$ of Section 2. To define an integrated information theory we must specify:

1. a class **Sys** of system types, closed under subsystems;
2. a definition of system cuts, under which **Sys** is closed;
3. a choice of weakly causal processes $\mathsf{caus}, \mathsf{eff}$ between the underlying objects M, P of each pair of subsystems $\underline{M}, \underline{P}$ via some state s, of any system $\underline{S}$.

More precisely, this provides the *data* of a generalised integrated information theory in the sense of (Kleiner and Tull, 2021). From this data we may now use the *IIT algorithm* from (Oizumi et al., 2014) to calculate the usual objects of interest in IIT.

6.1 The IIT Algorithm

We now briefly summarise this algorithm as treated in the general setting in (Kleiner and Tull, 2021), to which we refer for more details. Let us fix a 'current' state s of a system $\underline{S}$. Firstly, the level of *integration* of each value of the cause repertoire is defined by

$$\phi(\mathsf{caus}_s(M,P)) \;:=\; \min d(\mathsf{caus}_s(M,P)\;,\; \mathsf{caus}^{P_1,P_2}_{s,M_1,M_2}(M,P)) \qquad (7)$$

where the minima is taken over all pairs of decompositions of M, P which are not both trivial, i.e. equal to 1. [1] The integration level $\phi(\mathsf{eff}_s(M,P))$ is defined similarly in terms of eff.

For each choice of mechanism M, its *core cause* P^c and *core effect* P^e are the purviews P with maximal ϕ values for $\mathsf{caus}, \mathsf{eff}$ respectively. The minima of their corresponding ϕ values is then denoted by $\phi(M)$. We then associate to M and object called its *concept* $\mathbb{C}(M)$, essentially defined as the triple

$$(\mathsf{caus}_s(M,P^c), \mathsf{eff}_s(M,P^e), \phi(M))$$

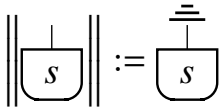
[1] When $\mathsf{caus}_s(M,P) = 0$ we alternatively set $\phi = 0$.

More precisely, in (Kleiner and Tull, 2021), $\mathbb{C}(M)$ is given by the pair of above repertoire values with each 'rescaled' by $\phi(M)$.

The tuple $\mathbb{Q}(s)$ of all these concepts, for varying M, is called the *Q-shape* $\mathbb{Q}(s)$ of the state s. The collection of all possible such tuples is denoted $\mathbb{E}(\underline{S})$. The level of integration of $\mathbb{Q}(s)$ is calculated similarly to (7) by considering all possible cuts of the system. The subsystem $\underline{M}$ of $\underline{S}$ whose Q-shape is itself found to be most integrated is called the *major complex*. Rescaling this Q-shape $\mathbb{Q}(\underline{M}, s|_M)$ according to its level of integration, and using an embedding $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$ we finally obtain a new element $\mathbb{E}(s) \in \mathbb{E}(\underline{S})$.

The claim of an IIT with regards to consciousness is that $\mathbb{E}(\underline{S})$ is the space of all possible conscious experiences of the system $\underline{S}$, and that $\mathbb{E}(s)$ is the particular experience attained when it is in the state s, with intensity $\Phi(s) := \| \mathbb{E}(s) \|$.

Remark 10 Let us make explicit how the specification of 1, 2, 3 above provides the data of an IIT in the sense of (Kleiner and Tull, 2021). The system class of the theory is **Sys**, and $\mathsf{caus}_s(M, P), \mathsf{eff}_s(M, P)$ and their decompositions are as outlined in Section 5.1. When **C** is probabilistic and has distances $d(a, b)$ defined for *arbitrary* states a, b of an object A, we may define the space of *proto-experiences* $\mathbb{PE}(\underline{S})$ of a system $\underline{S}$ to be simply its set of states, with



However, if d is only defined on causal states, as in classical IIT, to follow the algorithm from (Kleiner and Tull, 2021) one must instead set $\mathbb{PE}(\underline{S}) := \mathrm{St}_{\mathsf{c}}(S) \times \mathbb{R}^+$ as in Example 3 of (Kleiner and Tull, 2021). For either choice, for any subsystem $\underline{M}$ of $\underline{S}$ we obtain an embedding $\mathbb{PE}(\underline{M}) \hookrightarrow \mathbb{PE}(\underline{S})$ by composing alongside $\frac{1}{=}_{M^\perp}$, and this can be seen to provide a further embedding $\mathbb{E}(\underline{M}) \hookrightarrow \mathbb{E}(\underline{S})$.

7. Examples

Let us now meet several examples of IITs defined from process theories.

7.1 Generic IITs

Let $\mathbf{C}$ be any process theory coming with the features outlined in Section 2, including a dagger on processes. We define a generalised IIT denoted $\mathsf{IIT}(\mathbf{C})$ by taking as systems all tuples $\underline{S} = (S, \mathbb{D}, T)$ of an object S in $\mathbf{C}$ along with a causal process T and a decomposition set $\mathbb{D}$ induced by a single isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ in terms of elements S_i, as in Example 6. As before each partition of these elements gives a decomposition of S. We define system cuts to be symmetric as in (2) and the repertoires in the straightforward sense of (3).

Remark 11 We can extend this example in to ways. Firstly we may allow systems $\underline{S}$ to come with arbitrary finite decomposition sets $\mathbb{D}$ of S. Secondly, we may extend the definition to theories without daggers by instead simply requiring each system $\underline{S}$ to come with a process T^- describing 'reversed time evolution', and then define the cause repertoire by replacing $T^\dagger$ with T^-.

7.2 Classical IIT

The 'classical' IIT version 3.0 of Tononi and collaborators (Oizumi et al., 2014) is built on the process theory $\mathbf{Class}_\mathsf{m}$. As such a toy model of the theory is provided by $\mathsf{IIT}(\mathbf{Class}_\mathsf{m})$. However IIT 3.0 itself differs from this theory, using some more specific features of the process theories $\mathbf{Class}$ and $\mathbf{Class}_\mathsf{m}$ which we now describe.

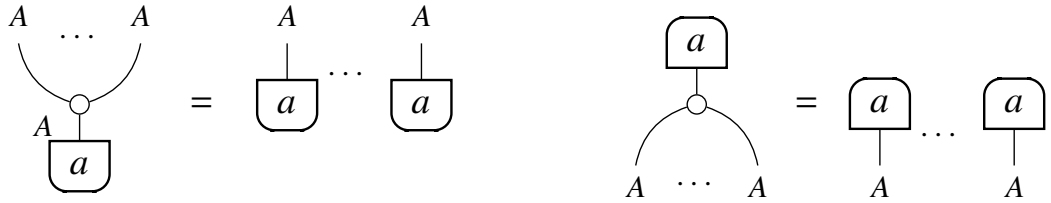
Firstly, note that in these classical process theories, for each object A, each element $a \in A$ corresponds to a unique state given by the point distribution at a, as well as a unique effect, namely the map sending a to 1 and all other elements of A to 0. We denote this state and effect both simply by a.[2]

Any process f from A to B is determined entirely by its compositions with these special states and effects since plugging in such a state a and effect b yields its value $f(a, b)$.

Another special feature of these classical process theories is that each object A comes with a distinguished *copying* process from A to $A \otimes \cdots \otimes A$,

[2]Typically these are the only kinds of 'state' considered, e.g. in (Oizumi et al., 2014) and even in our related article (Kleiner and Tull, 2021). In contrast here the term 'state' would include all distributions over A, i.e. all states of the process theory $\mathbf{Class}_\mathsf{m}$.

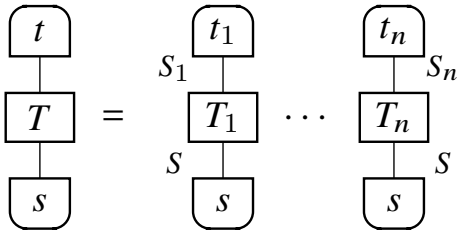
for any number of copies of A, as well as a *comparison* process in the opposite direction. We denote and define these respectively by the rules



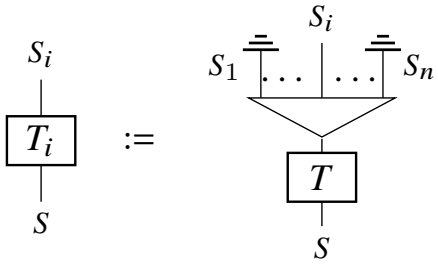
for all $a \in A$. Abstractly, these operations form a canonical commutative *Frobenius algebra* on each object, and there is no such canonical algebra on each object in **Quant** due to the *no-cloning* theorem (Coecke et al., 2013). We may now describe IIT 3.0 itself as follows.

7.2.1 Systems

In this theory systems are defined similarly to $\mathsf{IIT}(\mathbf{Class}_{\mathsf{m}})$, being given by a finite metric space S given as a product of elements $S \simeq \bigotimes_{i=1}^{n} S_i$, along with a causal (i.e. stochastic) evolution T on S. Additionally in (Oizumi et al., 2014) each evolution T is required to satisfy *conditional independence*, which states that for all $s, t \in S$, with $t = (t_1, \ldots, t_n)$ for some $t_i \in S_i$ we have

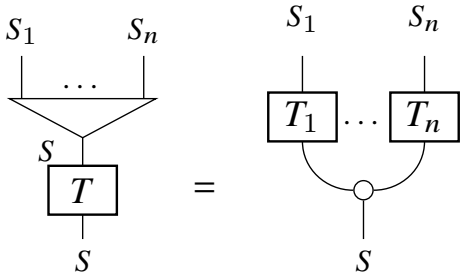


where for each element S_i we define the process T_i by



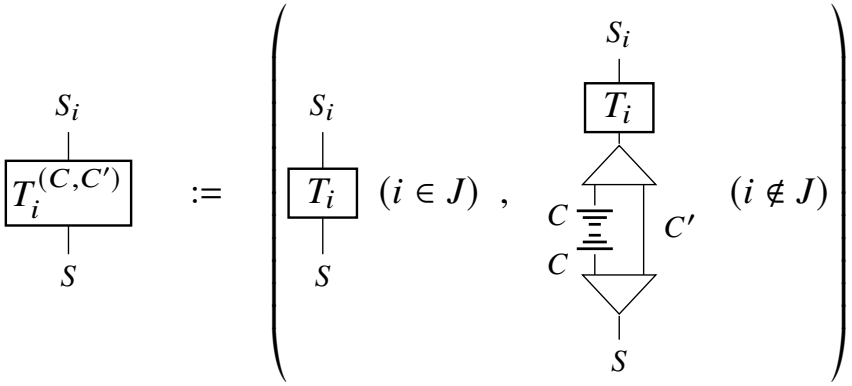
having depicted the isomorphism $S \simeq \bigotimes_{i=1}^{n} S_i$ by the triangle above. In other words, conditional independence states that the probabilities for the

next state of each element S_i are independent. Equivalently, T must satisfy



7.2.2 Cuts

Rather than our earlier symmetric cuts, the system cuts used in IIT 3.0 are *directional*. For any decomposition (C, C') of S with $C = \bigotimes_{j \in J} S_j$ for some subset of notes indexed by $J \subseteq \{1, \ldots, n\}$, we define the cut evolution $T^{(C,C')}$ using conditional independence by setting



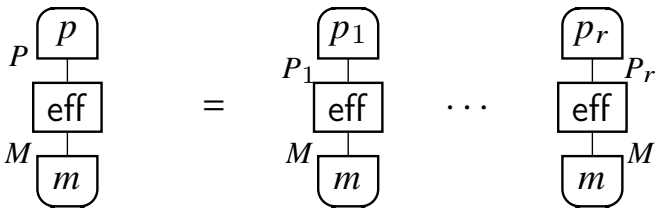
In other words, in the cut system all causal connections $C \to C'$ are replaced by noise, while all those into C remain intact.

7.2.3 Repertoires

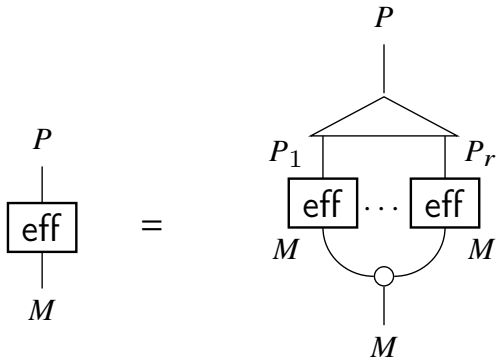
Let us now define the processes caus, eff between a pair of objects M and P, with $M = \bigotimes_{i=1}^{k} M_i$ and $P = \bigotimes_{j=1}^{r} P_j$ for some subsets $\{M_1, \ldots, M_k\}$ and $\{P_1, \ldots, P_r\}$ of elements of the system.

We begin with eff. When P is simply a single element P_j, eff is defined exactly as in (3). For more general P we define eff to again satisfy a form of

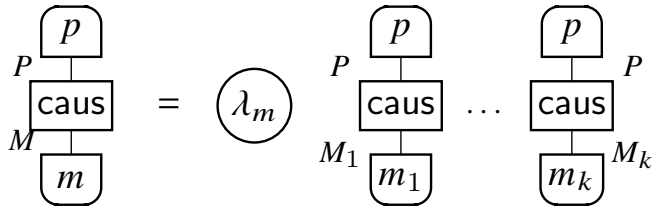
conditional independence, so that



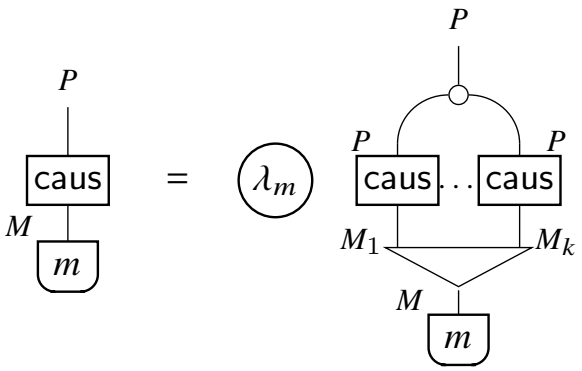
for all $m \in M, p = (p_1, \ldots, p_r) \in P$. Equivalently, we have that



In a similar fashion, whenever M is a single element M_i we define caus from M to P as in (4), while for more general M we require that



for all $m = (m_1, \ldots, m_k) \in M$ and $p \in P$, where λ_m is the normalisation scalar making caus $\circ\, m$ a causal state (probability distribution) if it is non-zero, or $\lambda_m = 0$ otherwise. Equivalently, this means that



for each $m \in M$. This concludes the data of classical IIT.

7.3 Quantum IIT

Zanardi, Tomka and Venuti have proposed a quantum extension of classical IIT (Zanardi et al., 2018). In fact it is comparatively much simpler to describe in our approach, being precisely the theory $\mathsf{IIT}(\mathbf{Quant})$.

Explicitly, systems in this theory are given by finite-dimensional complex Hilbert spaces $\mathcal{H}$ along with a given decomposition into elements $\mathcal{H} \simeq \bigotimes_{i=1}^{n} \mathcal{H}_i$ and a completely positive trace-preserving map T on $B(\mathcal{H})$. States and repertoire values are given by density matrices ρ. In this theory each Q-shape $\mathbb{Q}(\rho)$ may be encoded as a single positive semi-definite operator on the space $(\mathbb{C}^2)^{\otimes n} \otimes \mathbb{C}^2 \otimes \mathcal{H}$, as discussed in (Zanardi et al., 2018).

7.4 Quantum-Classical IIT

We may now define a version of *quantum-classical IIT* as $\mathsf{IIT}(\mathbf{CStar})$. This synthesizes quantum IIT with the toy version $\mathsf{IIT}(\mathbf{Class}_{\mathsf{m}})$ of classical IIT, containing both kinds of systems. In future it would be desirable to synthesise quantum IIT with IIT 3.0 proper. Since the latter relies on the presence of copying maps, this may be achievable using the more general notion of a *leak* on a C*-algebra (Selby and Coecke, 2017).

8. Outlook

In this article we have taken first steps to show how Integrated Information Theory, and its generalisations to other domains of physics, may be studied categorically. There are many avenues for future work.

Firstly, we have so far made no requirements on the cause and effect repertoire processes caus, eff. To be fit for their name these processes should be required to satisfy axioms which ensure they have a causal interpretation, ideally determining them uniquely within any given process theory. Monoidal categories provide a natural setting for the study of causality, a major contemporary topic in the foundations of physics (Kissinger and Uijlen, 2017).

At a higher level, it seems natural for the class of systems **Sys** of a generalised IIT to itself form a category. The theory itself should then give a functor into another category **Exp** of (spaces of) phenomenal experiences; a formalization of the latter is for example given in (Kleiner and Tull, 2021).

Making IIT functorial in this way will likely involve modifying it to be more natural from a categorical perspective. Developing a useful notion of integration applicable to any monoidal category may also help to resolve mathematical problems of the IIT algorithm, for example its relying on the unique existence of core purviews which are not guaranteed*.

References

Abramsky, Samson and Bob Coecke. 2004. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science, 2004.*, pages 415–425. IEEE.

Albantakis, Larissa, Arend Hintze, Christof Koch, Christoph Adami, and Giulio Tononi. 2014. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol* 10 (12): e1003966.

Albantakis, Larissa, William Marshall, Erik Hoel, and Giulio Tononi. 2017. What caused what? An irreducible account of actual causation. *arXiv:1708.06716* .

Alkire, Michael T, Anthony G Hudetz, and Giulio Tononi. 2008. Consciousness and anesthesia. *Science* 322 (5903): 876–880.

Barrett, Jonathan. 2007. Information processing in generalized probabilistic theories. *Physical Review A* 75 (3): 032304.

Beals, Richard, David H Krantz, and Amos Tversky. 1968. Foundations of multidimensional scaling. *Psychological review* 75 (2): 127.

Chiribella, Giulio, Giacomo Mauro D'Ariano, and Paolo Perinotti. 2010. Probabilistic theories with purification. *Physical Review A* 81 (6): 062348.

Clark, Austen. 1996. *Sensory Qualities*. Oxford Scholarship Online.

———. 2000. *A theory of sentience*. Clarendon press.

Coecke, Bob. 2014. Terminality implies non-signalling. *arXiv preprint arXiv:1405.3681* .

Coecke, Bob and Aleks Kissinger. 2017. *Picturing quantum processes*. Cambridge University Press.

*ACKNOWLEDGEMENTS: We would like to thank the organizers and participants of the *Workshop on Information Theory and Consciousness* at the Centre for Mathematical Sciences of the University of Cambridge, of the *Modelling Consciousness Workshop* in Dorfgastein and of the *Models of Consciousness Conference* at the Mathematical Institute of the University of Oxford for discussions on this topic. Much of this work was carried out while Sean Tull was under the support of an EPSRC Doctoral Prize at the University of Oxford, from November 2018 to July 2019, and while Johannes Kleiner was under the support of postdoctoral funding at the Institute for Theoretical Physics of the Leibniz University of Hanover. We would like to thank both institutions.

Coecke, Bob and Eric Oliver Paquette. 2010. Categories for the practising physicist. In *New structures for physics*, pages 173–286. Springer.

Coecke, Bob, Dusko Pavlovic, and Jamie Vicary. 2013. A new description of orthogonal bases. *Mathematical Structures in Computer Science* 23 (3): 555–567.

Ehresmann, Andrée C. 2012. Mens: from neurons to higher mental processes up to consciousness. In *Integral Biomathics*, pages 29–30. Springer.

Fekete, Tomer and Shimon Edelman. 2011. Towards a computational theory of experience. *Consciousness and cognition* 20 (3): 807–827.

Fink, Sascha Benjamin, Wanja Wiese, and Jennifer Michelle Windt. 2018. *Philosophical and Ethical Aspects of a Science of Consciousness and the Self*. Frontiers in Psychology.

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

Kissinger, Aleks and Sander Uijlen. 2017. A categorical semantics for causal structure. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12. IEEE.

Kleiner, Johannes and Sean Tull. 2021. The Mathematical Structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics* .

Marshall, William, Jaime Gomez-Ramirez, and Giulio Tononi. 2016. Integrated information and state differentiation. *Frontiers in psychology* 7: 926.

Marshall, William, Hyunju Kim, Sara I Walker, Giulio Tononi, and Larissa Albantakis. 2017. How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375 (2109): 20160358.

McDermott, Drew. 2007. Artificial intelligence and consciousness. *The Cambridge handbook of consciousness* pages 117–150.

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

Michel, Matthias, Diane Beck, Ned Block, Hal Blumenfeld, Richard Brown, David Carmel, Marisa Carrasco, Mazviita Chirimuuta, Marvin Chun, Axel Cleeremans, et al. 2019. Opportunities and challenges for a maturing science of consciousness. *Nature human behaviour* 3 (2): 104–107.

Northoff, Georg, Naotsugu Tsuchiya, and Hayato Saigo. 2019. Mathematics and the brain: A category theoretical approach to go beyond the neural correlates of consciousness. *Entropy* 21 (12): 1234.

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. 2014. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology* 10 (5): e1003588.

Selby, John and Bob Coecke. 2017. Leaks: quantum, classical, intermediate and

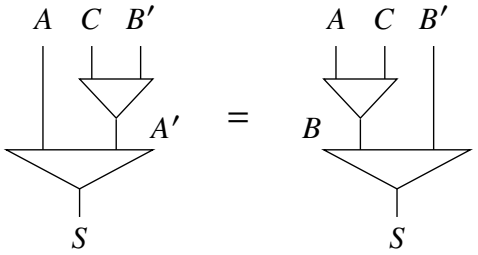
more. *Entropy* 19 (4): 174.
Selinger, Peter. 2007. Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical computer science* 170: 139–163.
———. 2011. A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer.
Tononi, Giulio. 2004. An information integration theory of consciousness. *BMC neuroscience* 5 (1): 42.
———. 2008. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin* 215 (3): 216–242.
———. 2015. Integrated information theory. *Scholarpedia* 10 (1): 4164.
Tsuchiya, Naotsugu, Shigeru Taguchi, and Hayato Saigo. 2016. Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience research* 107: 1–7.
Wiese, Wanja and Karl Friston. 2020. The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation .
Zanardi, Paolo, Michael Tomka, and Lorenzo Campos Venuti. 2018. Quantum integrated information theory. *arXiv preprint arXiv:1806.01421* .

A. Decompositions and Integration

Here we briefly mention a few further results about decompositions of objects in process theories; we leave a detailed study of their properties to future work.

Our earlier definition of $\mathbb{D}|_A$ was based on an idea of one decomposition as being 'contained in' another. Let us make this precise.

Definition 12 *Let S be an object in a process theory and (A, A'), (B, B') two decompositions. We write that $(A, A') \preceq (B, B')$ whenever there exists an object C and decompositions (A, C) of B and (B', C) of A' such that*

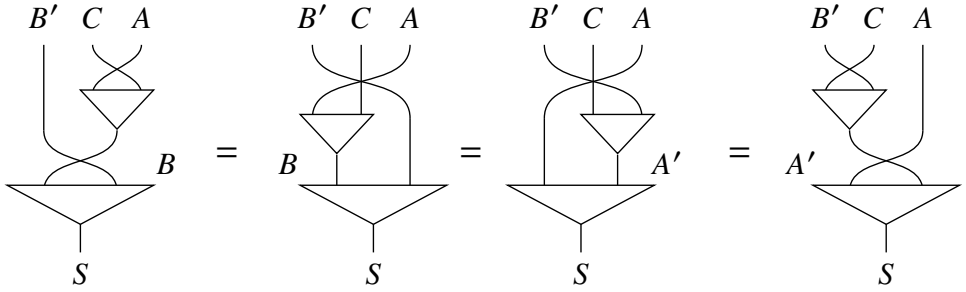


(8)

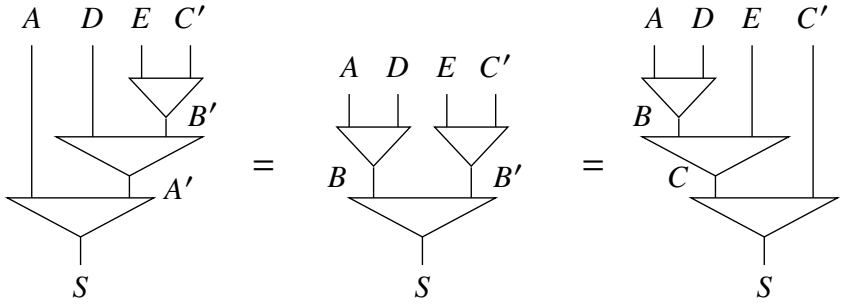
Intuitively, this states that A is contained in B (as is B' within A') in a way compatible with these decompositions.

Lemma 13 *Let S be an object in a process theory. Then $\preceq$ forms a pre-order on the set of decompositions of S, with top element 1 and bottom element 0, and $(-)^\perp$ as an involution.*

Proof. We always have $(A, A') \preceq (A, A')$ by taking $C = I$ and using the decompositions 1 and 0 on A in (8). Similarly $(A, A') \preceq 1$ by taking $C = A'$. To see that $(-)^\perp$ is an involution, suppose that $(A, A') \preceq (B, B')$ as above. Then we have $(B, B')^\perp \preceq (A, A')^\perp$ since



Hence we always have $0 = 1^\perp \preceq (A, A')$ for all (A, A'). For transitivity, note that whenever $(A, A') \preceq (B, B') \preceq (C, C')$ via some respective objects D, E then we have

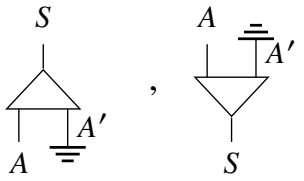


so that $(A, A') \preceq (C, C')$ via the above decompositions $(D \otimes E, C')$ of A' and $(A, D \otimes E)$ of C. ■

Recall that in any category, a *sub-object* of an object A is an (isomorphism class of a) monomorphism $m\colon M \to A$. It is *split* when $e \circ m = \mathrm{id}_M$ for some e. The sub-objects of A form a partial order $\mathrm{Sub}(A)$.

Lemma 14 *In any process theory with* $\overline{\top}, \overline{\bot}$*, for any object* S*:*

1. *Any decomposition* (A, A') *of* S *makes* A *a split sub-object of* S *via*

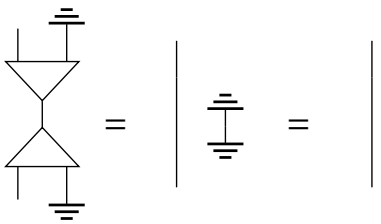


(9)

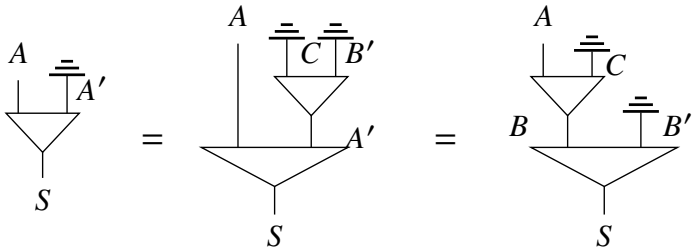
   Moreover if $(A, A') \preceq (B, B')$ *then* $A \leq B$ *in* $\mathrm{Sub}(S)$*.*

2. $\preceq$ *restricts to a partial order* $\leq$ *on* $\mathbb{D}(S)$*, again with top element* 1*, bottom* 0 *and involution* $(-)^{\perp}$*.*

Proof. 1: We have

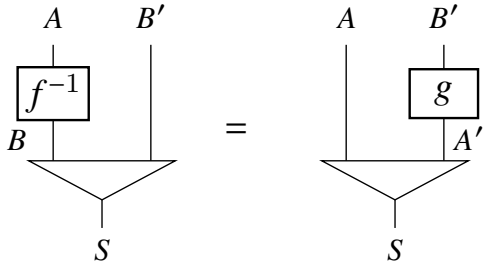


If $(A, A') \preceq (B, B')$ then the splitting for A factors over that for B since:



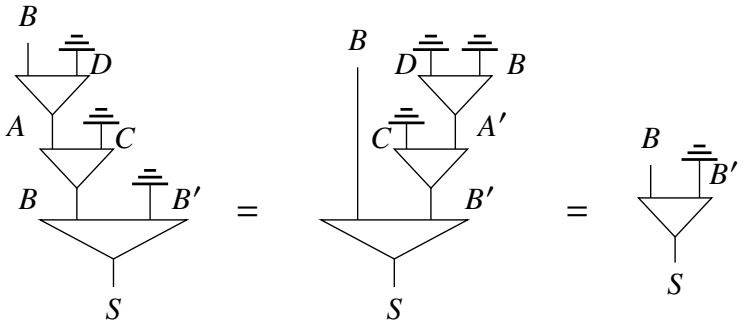
It follows that $A \leq B$ in $\mathrm{Sub}(S)$.

2: We need to show that any two decompositions (A, A') and (B, B') are equivalent under $\preceq$ precisely when they are equivalent in the sense of (1). Firstly, if there exists causal and co-causal isomorphisms f, g making (1) hold, then we have

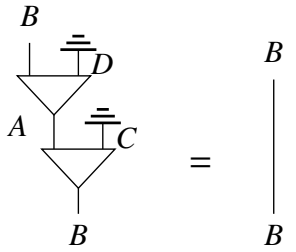


Viewing f^{-1} and g as decompositions (A, I) of B and (I, B') of A', respectively, this gives that $(B, B') \preceq (A, A')$. Then $(A, A') \preceq (B, B')$ holds similarly.

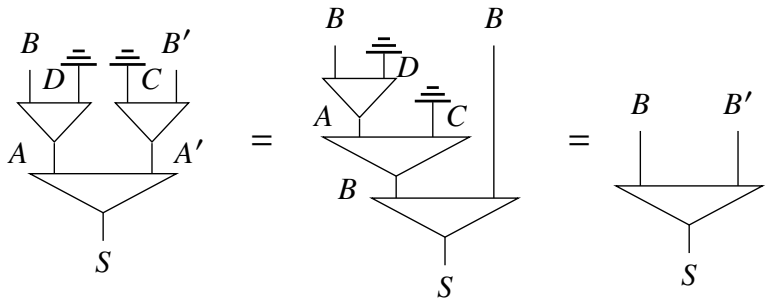
Conversely, if $(A, A') \preceq (B, B') \preceq (A, A')$, via respective objects C, D then



Since the right-hand map is an epimorphism by the first part, this gives that



Dually, composing in the other order gives the identity on A, making these causal and co-causal isomorphisms $A \simeq B$. Similarly we obtain such isomorphisms $A' \simeq B$'. Then we have



as required. Now 2 follows since any pre-order restricts to a partial order on its set of equivalence classes, and so $\preceq$ becomes a partial order $\leq$ on $\mathbb{D}(S)$. It is easy to see that the earlier properties of $1, 0, (-)^{\perp}$ carry over to $\leq$. ■

A.1 Integration

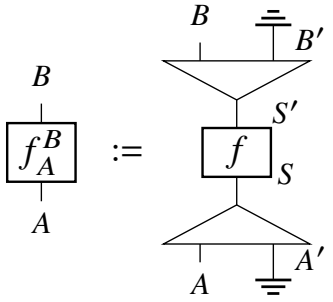
Let us briefly allude to how integration may generally be studied and quantified using decomposition sets.

Suppose we have objects S, S' with given decomposition sets $\mathbb{D}, \mathbb{D}'$ and for each $(A, A') \in \mathbb{D}$ and $(B, B') \in \mathbb{D}'$ a process f_A^B from A to B. We denote $f_S^{S'}$ simply by f. Whenever we have a given distance function d on the set of processes from S to S', we may define the level of *integration* of the family $(f_A^B)_{A,B}$ as

$$\phi(f) := \min_{\mathbb{D}\times\mathbb{D}'} d\left(\boxed{f}_{S}^{S'} \,,\; \boxed{f_A^B}_{A}^{B} \; \boxed{f_{A'}^{B'}}_{A'}^{B'} \right)$$

where we exclude the top element $(1, 1)$ of $\mathbb{D} \times \mathbb{D}'$ in the minimisation.

Example 15 Given any process f from S to S' we may define such a family $(f_A^B)_{A,B}$ with $f_S^{S'} = f$ by setting



Example 16 Our earlier description of the IIT algorithm precisely includes evaluating the integration level of each of the families of processes $(\mathsf{caus})_{M,P}$ and $(\mathsf{eff})_{M,P}$ using the state-dependent distance

$$d_m\left(\boxed{f}_{M}^{P} \,,\; \boxed{g}_{M}^{P} \right) := d\left(\boxed{\boxed{f}\,m}^{P} \,,\; \boxed{\boxed{g}\,m}^{P} \right)$$

where $m = s|_M$ and d is the distance on $\mathrm{St}(S)$.

Neuroscience of Consciousness, 2021, 7(1): niab001

doi: 10.1093/nc/niab001
Research article

Falsification and consciousness

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

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

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

Abstract

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

Keywords: consciousness; theories and models

Introduction

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

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

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

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

Received: 9 July 2020; **Revised:** 23 November 2020. **Accepted:** 5 January 2021

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

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

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

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

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

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

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

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

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

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

Formal Description of Testing Theories

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

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

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

Experiments

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

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

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

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

Note that obs describes the experiment, the choice of observables, and all conditions during an experiment that generates the dataset o necessary to apply the theory, which may differ from theory to theory, such as interventions in the case of IIT. In all realistic cases, the correspondence obs is likely quite complicated since it describes the whole experimental setup. For our argument, it simply suffices that this mapping exists, even if it is not known in detail.

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

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

Predictions

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

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

$$\mathrm{pred} : \mathcal{O} \twoheadrightarrow E\,.$$

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

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

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

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

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

Inferences

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

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

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

$$\mathrm{inf} : \mathcal{O} \to \mathrm{E}\,.$$

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

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

Falsification

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

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

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

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

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

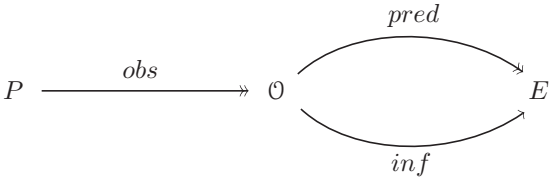


Figure 2. Two maps are necessary for a full experimental setup, one that describes a theory's predictions about experience (*pred*), another that describes the experimenter's inference about it (*inf*). Both map from a dataset $o \in \mathcal{O}$ collected in an experimental trail to some subset of experiences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

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

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

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

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

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

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

Summary

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

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

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

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

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

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

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

The Substitution Argument

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

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

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

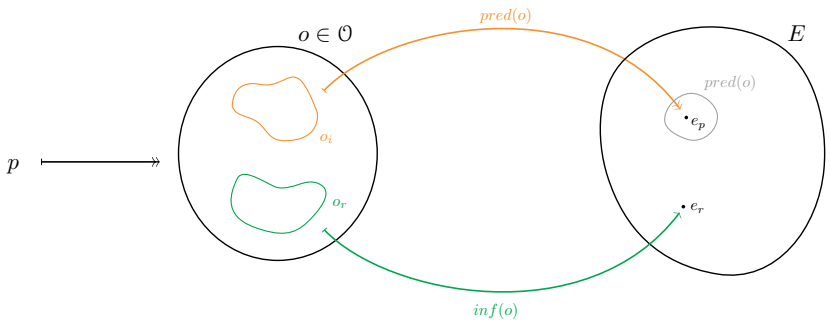


Fig. 3. This figure represents the same setup as Fig. 2. The left circle depicts one single dataset o. o_i (orange) is the part of the dataset used for prediction. o_r (green) is the part of the dataset used for inferring the state of experience. Usually the green area comprises verbal reports or button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

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

Substitutions

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

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

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

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

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

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

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

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

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

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

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

Substitutions imply falsifications

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

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

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

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

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

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

Universal substitutions imply complete falsification

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

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

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

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

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

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

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

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

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

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

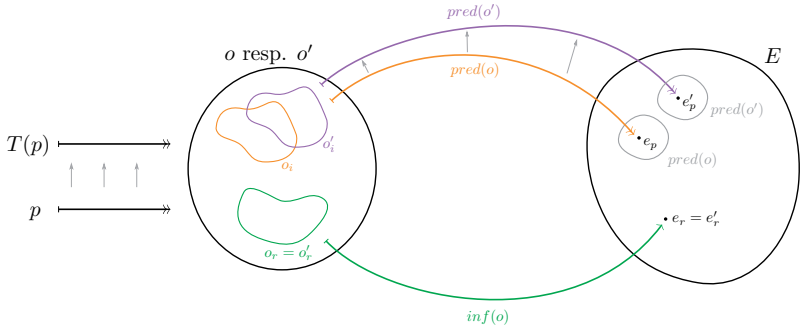


Figure 4. This picture illustrates substitutions. Assume that some dataset o with inference content o_r is given. A substitution is a transformation S of physical systems which leaves the inference content o_r invariant but which changes the result of the prediction process. Thus whereas p and $S(p)$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

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

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

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

When does a universal substitution exist?

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

Theories need to be minimally informative

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

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

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

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

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

$$\mathrm{pred}(\overline{o}) \cap \mathrm{pred}(\overline{o}') = \varnothing\,. \tag{8}$$

Inference and prediction data are independent

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

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

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

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

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

Universal substitutions exist

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

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

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

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

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

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

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

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

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

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

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

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

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

Examples of data independence

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

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

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

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

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

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

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

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

Inference and Prediction Data Are Strictly Dependent

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Our second main theorem is the following.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Objections

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

Restricting inferences to humans only

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

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

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

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

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

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

Reductio ad absurdum

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

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

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

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

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

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

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

Theories could be based on phenomenology

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

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

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

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

Rejecting falsifiability

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

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

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

Conclusion

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

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

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

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

Lenient dependency

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

A priori, in the space of possible relationships between inference and prediction data, there seems to be room for relationships that are neither independent (The substitution argument section) nor strictly dependent (Inference and prediction data are strictly dependent section). We define this relationships of this kind as cases of *lenient dependency*. No current theory or testing paradigm that we know of satisfies this definition. Yet cases of lenient dependency cannot be excluded to exist. Such cases would technically not be beholden to either Theorem 3.10 or Theorem 4.3.

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

Physics is not causally closed

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

Supplementary Materials

(A) Weak Independence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(B) Inverse Predictions

When defining falsification, we have considered predictions that take as input data about the physical configuration of a system and yield as output a state of consciousness. An alternative would be to consider the inverse procedure: a prediction which takes as input a reported stated of consciousness and yields as output some constraint on the physical configuration of the system that is having the conscious experience. In this section, we discuss the second case in detail.

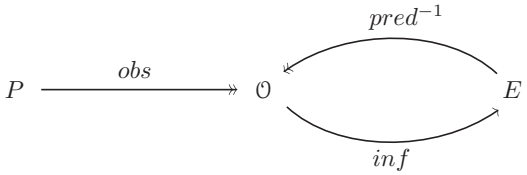


Fig. 5. The case of an inverse prediction. Rather than comparing the inferred and predicted state of consciousness, one predicts the physical configuration of a system based on the system's report and compares this with measurement results.

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

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

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

Accordingly, the class of all prediction data which is compatible with the inferred experience $\mathrm{inf}(o)$ is

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

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

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

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

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

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

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

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

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

Acknowledgments

We would like to thank David Chalmers, Ned Block, and the participants of the NYU philosophy of mind discussion group for valuable comments and discussion. Thanks also to Ryota Kanai, Jake Hanson, Stephan Sellmaier, Timo Freiesleben, Mark Wulff Carstensen, and Sofiia Rappe for feedback on early versions of the manuscript.

Conflict of interest statement. None declared.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Doerig A, Schurger A, Hess K, *et al.* The unfolding argument: why iit and other causal structure theories cannot explain consciousness. *Conscious Cogn* 2019;**72**:49–59.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

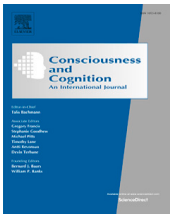
Mather JA. Cephalopod consciousness: behavioural evidence. *Conscious Cogn* 2008;**17**:37–48.
Mediano P, Seth A, Barrett A. Measuring integrated information: comparison of candidate measures in theory and simulation. *Entropy* 2019;**21**:17.
Napolitani M, Bodart O, Canali P, *et al*. Transcranial magnetic stimulation combined with high-density EEG in altered states of consciousness. *Brain Injury* 2014;**28**:1180–9.
Negro N. Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenol Cogn Sci* 2020;19:979–96.
Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 2014;**10**:e1003588.
Popper K. *The Logic of Scientific Discovery*. New York: Harper & Row, 1968.
Putnam H. Minds and machines. In *Dimensions of Mind*, ed. S. Hook, New York: New York University Press, 1960, pp. 57–80.
Radnitzky G. Review: Refined falsificationism meets the challenge from the relativist philosophy of science. Br J Philos Sci 1991;**42**:273–284.
Reardon S. Rival theories face off over brain's source of consciousness. Science 2019;**366**:293–293.
Richards BA, Lillicrap TP, Beaudoin P, *et al*. A deep learning framework for neuroscience. *Nat Neurosci* 2019;**22**:1761–70.
Rosenthal DM. How many kinds of consciousness? *Conscious Cogn* 2002;**11**:653–65.
Rounis E, Maniscalco B, Rothwell JC, *et al*. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* 2010;**1**:165–75.
Schäfer AM, Zimmermann HG. Recurrent neural networks are universal approximators. In: *International journal of neural systems*. Springer, 2007;**17**:253–63.
Sergent C, Dehaene S. Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *J Physiol Paris* 2004;**98**:374–84.
Seth AK. Models of consciousness. *Scholarpedia* 2007;**2**:1328.
Seth AK. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn Neurosci* 2014;**5**:97–118.
Skinner BF. The behavior of organisms: an experimental analysis. *Appleton-Century*, Cambridge, Massachusetts: B.F. Skinner Foundation. 1938.
Solomonoff RJ. A formal theory of inductive inference. Part I. *Inform Control* 1964;**7**:1–22.
Tononi G. An information integration theory of consciousness. *BMC Neurosci* 2004;**5**:42.
Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 2008;**215**:216–42.
Giulio T. Why Scott should stare at a blank wall and reconsider (or, the conscious grid). In: *Shtetl-Optimized: The Blog of Scott Aaronson*. Available online: http://www. scottaaronson.com/blog, 2014.
Tsuchiya N, Wilke M, Frässle S, *et al*. No-report paradigms: extracting the true neural correlates of consciousness. *Trends Cogn Sci* 2015;**19**:757–70.
Tsuchiya N, Taguchi S, Saigo H. Using category theory to assess the relationship between consciousness and integrated information theory. *Neurosci Res* 2016;**107**:1–7.
Tsuchiya N, Andrillon T, Haun A. A reply to "the unfolding argument": beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. PsyArXiv, 2019.
Turing AM. Computing machinery and intelligence. Mind 1950; **59**:433–60.
Turing AM. On computable numbers, with an application to the entscheidungsproblem. *Proc Lond Math Soc* 1937;**2**: 230–65.
Wenzel M, Han S, Smith EH, *et al*. Reduced repertoire of cortical microstates and neuronal ensembles in medically induced loss of consciousness. *Cell Syst* 2019;**8**:467–74.
Wolfram S. Cellular automata as models of complexity. *Nature* 1984;**311**:419.
Wu T, Tegmark M. Toward an artificial intelligence physicist for unsupervised learning. *Phys Rev E* 2019;**100**:033311.

Consciousness and Cognition 119 (2024) 103653

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/yccog



Full Length Article

Towards a structural turn in consciousness science

Johannes Kleiner [a,b,c,d,*]

[a] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 München, Germany
[b] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany
[c] Institute for Psychology, University of Bamberg, Markusplatz 3, 96047 Bamberg, Germany
[d] Association for Mathematical Consciousness Science, Geschwister-Scholl-Platz 1, 80539 München, Germany

ARTICLE INFO

Keywords:
Quality space
Qualia space
Phenomenal space
Theory of consciousness
Structuralism
Structure-preserving mapping
Isomorphism

ABSTRACT

Recent activities in virtually all fields engaged in consciousness studies indicate early signs of a structural turn, where verbal descriptions or simple formalisations of conscious experiences are replaced by structural tools, most notably mathematical spaces. My goal here is to offer three comments that, in my opinion, are essential to avoid misunderstandings in these developments early on. These comments concern metaphysical premises of structural approaches, the viability of structure-preserving mappings, and the question of what a structure of conscious experience is in the first place. I will also explain what, in my opinion, are the great promises of structural methodologies and how they might impact consciousness science at large.

1. Introduction

So far, the scientific study of consciousness has mainly employed verbal and linguistic tools, as well as simple formalisations thereof, to describe conscious experiences. Typical examples are the distinction between 'being conscious' and 'not being conscious', between whether a subject is 'perceiving a stimulus consciously' or not, between whether a subject is 'experiencing a particular quale' rather than another, or more generally any account of whether some X is part of the phenomenal character of a subject's experience at some point of time. Formalisations of these verbal descriptions mostly make use of set theory, examples being sets of states of consciousness of a subject and simple binary classifications, or of real numbers, for example to model 'how conscious' a system is. There are sophisticated mathematical techniques in the field, but to a large extent they only concern the statistical analysis of empirical data, and the formulation of a theory of consciousness itself—but not the description of conscious experiences which underlies the data collection or modelling effort.

Much like words shape thoughts, descriptions shape science. In the case of consciousness studies, the descriptions that were available so far have fed into theories of consciousness, have determined what can be inferred about the state of consciousness of a subject, and have guided ways of conceptualising the problem under investigation.

They have, for example, led to a number of theories that explain what it takes for a single stimulus or a single piece of information to be consciously experienced, but which remain silent or vague on how the phenomenal character as a whole is determined. They have led to measures of consciousness which are specifically tailored to find out whether a single stimulus or single quality is experienced consciously (Irvine, 2013), but are not meant to infer phenomenal character beyond this. And to some extent, at least, they have privileged research programmes which search for either-or conditions related to consciousness, such as arguably the

* Correspondence to: Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany.
E-mail address: johannes.kleiner@lmu.de.

https://doi.org/10.1016/j.concog.2024.103653
Received 6 October 2023; Received in revised form 22 January 2024; Accepted 30 January 2024

Available online 28 February 2024

search for Neural Correlates of Consciousness (NCCs) that is largely predicated on a conception of having "any one specific conscious percept" (Koch et al., 2016).

Because verbal descriptions only parse part of the phenomenal character of an experience, part of what it is like for an organism to live through a particular moment, it is no surprise that means to go beyond these simple descriptions are highly sought after.

In recent years, the idea of using mathematical spaces, or mathematical structure more generally,[1] to go beyond verbal descriptions and simple formalisations have started to sprout in virtually every discipline involved in the scientific quest to understand consciousness. Following rich developments in psychophysics over more than a century (Pashler & Wixted, 2004), and pioneering work by Austen Clark (Clark, 1993) and David Rosenthal (Rosenthal, 1991) in consciousness science, mathematical spaces are now applied in philosophy, (Clark, 2000, Coninx, 2022, Fortier-Davy & Millière, 2020, Gert, 2017, Lee, 2021, 2022, Rosenthal, 2010, 2015, 2016, Fink et al., 2021, Lyre, 2022, Kob, 2023, Renero, 2014, Prentner, 2019, Yoshimi, 2007, Chalmers & McQueen, 2022, Silva, 2023, Atmanspacher, 2020), neuroscience (Tononi, 2015, Tallon-Baudry, 2022, Zaidi et al., 2013, Lau et al., 2022, Malach, 2021, Haun & Tononi, 2019, Oizumi et al., 2014, Hebart et al., 2020, Josephs et al., 2023, Tsuchiya et al., 2023, Zeleznikow-Johnston et al., 2023, Haynes, 2009, Michel, In press), cognitive science (Hoffman et al., 2023, Rudrauf et al., 2017, Hoffman & Prakash, 2014, O'Brien & Opie, 1999), psychology (Klincewicz, 2011, Kostic, 2012, Young et al., 2014) and mathematical consciousness science (Grindrod, 2018, Kleiner, 2020b, Stanley, 1999, Resende, 2022, Mason, 2013, 2021, Signorelli & Wang & Coecke, 2021, Tsuchiya et al., 2016, Tsuchiya & Saigo, 2021, Tsuchiya et al., 2022, Kleiner, 2020a, Kleiner & Hoel, 2021, Kleiner & Ludwig, 2023). They are known under various names, including quality spaces (Clark, 1993, Rosenthal, 2015), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021, Kleiner & Tull, 2021, Rosenthal, 2010), qualia structure (Kawakita & Zeleznikow-Johnston & Tsuchiya, et al., 2023, Kawakita & Zeleznikow-Johnston & Takeda, et al., 2023, Tsuchiya et al., 2022), Q-spaces (Chalmers & McQueen, 2022, Lyre, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink et al., 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). A first formalised theory of consciousness to make use of mathematical spaces was Integrated Information Theory (IIT) 2.0 (Tononi, 2008); more recent versions expand and refine the idea (Oizumi et al., 2014, Albantakis et al., 2023).

What unites all of these proposals is the hope that the mathematical structures they propose are useful to describe the phenomenal character of an experience more comprehensively, more precisely, or more holistically than verbal descriptions or simple formalisations allow, and that mathematical structures can cope both with the apparent richness and with the many details that make up experiences. If this hope turns out true, it has far-reaching implications on how to study, measure, and think about consciousness.

My goal here is to offer three comments which I think are important to keep in mind when applying structural ideas in theory and experimental practice, so as to avoid misconceptions or misunderstanding early on. I hope that my comments are helpful for those working on structural ideas as well as those observing these developments with a degree of scepticism.

2. Three promises of a structural turn

Before offering my comments below, I will briefly sketch the implications and limitations that structural methodologies may have for consciousness science. This might be of interest to those who have not engaged with this research before, and allows me to illustrate what I think are some of the great promises of a structuralist turn.

2.1. Theories of consciousness

We currently have at least 39 theories of consciousness,[2] with new theories being proposed on a regular basis, albeit without much general attention. The reason for that, I contend, is that as far as theoretical work is concerned, it is actually very easy to come up with theories of consciousness of the type we have today.

The majority of contemporary theories of consciousness aim to explain whether a system's state, a stimulus, a piece of information, or a representation is consciously experienced, or not. That is, they target a *binary classification* between states, signals, stimuli or

[1] The term *mathematical structure*, which I will explain in detail Section 3 below, is more general than the term *mathematical space*. That is, every mathematical space is a mathematical structure, but there are also mathematical structures which are not mathematical spaces, either because they only comprise individuals (so do not satisfy the intuition that a space is about many individuals), or because their structure is more complex than one would typically take a space to be. The question of which mathematical structures to call mathematical spaces is a matter of convention, which is why there is no definition of a general concept of mathematical space in mathematical logic.

[2] An unpublished list compiled by Dr. Jonathan Mason on behalf of the *Association for Mathematical Consciousness Science* (AMCS) and the *Oxford Mathematics of Consciousness and Applications Network* (OMCAN) comprises the following theories of consciousness in the peer-reviewed literature: Activation/Information/Mode-Synthesis Hypothesis, Adaptive Resonance Theory, Attention Schema Theory, Centrencephalic Proposal, Conscious Agent Networks, Conscious Turing Machine, Consciousness Electromagnetic Information Field Theory, Consciousness State Space Model, Cross-Order Integration Theory, Dendrite/Apical Dendrite Theory, Dynamical Core Theory, Electromagnetic Field Hypothesis, Enactive and Radical Embodiment, Expected Float Entropy Minimisation, First Order Representational Theory, Free Energy Principle Projective Consciousness Model, Global Neuronal Workspace Theory, Global Workspace Theory, Higher-Order Thought Theory, Integrated Information Theory, Integrated World Modeling Theory, Layered Reference Model of the Brain, Memory Consciousness and Temporality Theory, Mesocircuit Hypothesis, Multiple Draft Model, Network Inhibition Hypothesis, Neural Darwinism Theory, Orchestrated Objective Reduction, Passive Frame Theory, Predictive Processing and Interoception, Proto-Consciousness Induced Quantum Collapse, Psychological Theory of Consciousness, Radical Plasticity Thesis, Recurrent Processing Theory, Self Comes to Mind Theory, Semantic Pointer Competition Theory, Single Particle Consciousness Hypothesis, Temporo-Spatial Theory of Consciousness, Thalamo-Cortical Loops and Sensorimotor Couplings. This list might not be complete, and some of the theories might point to similar or analogous theoretical constructs.

representations. The simple verbal distinctions mentioned in the introduction—a system 'being conscious' or not, 'perceiving a stimulus consciously' or not, 'experiencing a particular quale' or not—are all examples of such binary classifications.

Formulating theories of consciousness that target binary classification is relatively straightforward, as far as theoretical work is concerned. This is because devising a $\{0,1\}$ classification only requires identifying some property, function, or dynamical mode of a brain mechanism. All configurations that exhibit this property, function or dynamical mode are mapped to 1, while all which do not are mapped to 0. And within non-structural approaches, nothing technical prohibits one from postulating that the 1 cases correspond to conscious experience of a stimulus, state, piece of information or representation, while the 0 cases correspond to unconscious experience thereof. The empirical or conceptual validity of such a choice is an important question, yet from a technical standpoint, formulating theories that target these distinctions is straightforward.

It is much more difficult to come up with a well-formed hypothesis that relates to a mathematical space or mathematical structure. That is because a mathematical space or mathematical structure has two parts. On the one hand, it contains a set of points. On the other hand, it contains relations or functions that express connections between the points, for example an order relation or a metric function. Therefore, there is much more information to provide when specifying how a space or structure relates to a brain mechanism, or a physical system more generally. Furthermore, virtually every mathematical object comes with a set of axioms that parts of the object have to satisfy. So not only is more information needed, but this information may also have to satisfy constraints to provide a legitimate definition. This is why defining a space or structure is much more of a challenge than finding a binary classification.

The task is more difficult even if the space or structure that a theory is to provide has a specific, theory-independent form. That is the case if the theory has to account for phenomenal structure that has independent justification or independent motivation, for example from psychophysical experiments. This difficulty is illustrated by the fact that we do not, at present, have a theory of consciousness that targets the mathematical structures that have been proposed to account for conscious experiences on independent grounds. To the best of my knowledge, there are only two theories that define phenomenal spaces: Integrated Information Theory (IIT) (Albantakis et al., 2023) and Expected Float Entropy Minimisation Theory (EFE) (Mason, 2021). While both theories represent significant advances, establishing a link to existing phenomenal spaces (cf. Section 5) remains a next-level challenge.[3]

As formulating theories that account for phenomenal structure in addition to non-structural explananda necessitates meeting more constraints than formulating non-structural theories, structural theories are likely to be more predictive than their non-structural counterparts. Furthermore, because the phenomenal structure is an integral aspect of phenomenal character, a theory that accounts for phenomenal structure in addition to non-structural explananda has a broader explanatory scope than one that focuses solely on the conscious-unconscious distinction. Therefore, a structural turn might deliver more explanatory and more predictive theories of consciousness. This is the first major implication I can see of structural approaches in consciousness science.[4]

Structural methodologies might inspire, and be inspired by, novel theoretical ideas that derive from any of the existing theories of consciousness, or from their combination. Proposals like the Conscious Turing Machine (Blum & Blum, 2022) or Integrated World Modelling Theory (Safron, 2022) that combine features of existing theories of consciousness (such as, for example, Integrated Information Theory, Global Neuronal Workspace Theory, and Free Energy Principle based proposals) could be particularly interesting in this regard.

2.2. Experimental investigations

A shift towards structural methodologies could also have significant implications for experimental research. One immediate implication follows from the previous section, i.e., from the transformative effect that structural methodologies could have on theories of consciousness. If structural theories of consciousness would indeed be more predictive than the non-structural theories we have today, then they might be easier to test than the theories we have today,[5] and the new predictions about structural facts might offer new avenues for experimental investigation.[6]

But structural thinking could also yield new experimental tools and methodologies that are separate from theoretical advancements. For instance, under certain conditions, structural approaches offer an entirely new methodology for measuring NCCs (Fink et al., 2021). This methodology could potentially address some of the foundational challenges in existing methodologies, such as the co-activation of cognitive processing centres causally downstream of the core NCC, and might not require traditional methods to assess a subject's state of consciousness. I discuss and criticise the key assumption that enables this methodology—the assumption of a structure-preserving mapping between phenomenal and neuronal structures—in Section 4 below. Nevertheless, even if this

[3] Proponents of both theories are fully aware of this task, and IIT has made a first step in this direction in Haun and Tononi (2019). In addition to accounting for phenomenal structure that has independent justification, there are other tasks and challenges that structural theories have to meet and resolve. For example, an anonymous reviewer has kindly pointed me to the fact that according to IIT, richly structured experience can be entailed by static systems without dynamics, which might pose an empirical or conceptual challenge for IIT.

[4] In saying this, I do not intend to diminish the value of 'binary' theories of consciousness. They are an integral part of consciousness science and encapsulate a substantial body of evidence. On my view, they need to be extended so as to address phenomenal character more holistically as well. Whether this should be done on a case-by-case basis, or whether there might be a theory of qualitative character that can serve for a larger number of binary theories, is not something that needs to be decided in advance.

[5] Lukas Kob made this point for *structuralist* approaches during a wonderful talk at the recent *Structuralism in Consciousness Studies* workshop at the Charité Berlin, though my comment here concerns the wider scope of *structural* approaches, cf. Section 3 and Fig. 2 for more on that distinction.

[6] Speculating wildly, one might hope that if theories of consciousness could account for *theory-independent* phenomenal spaces, this could help to mitigate the problem that empirical tests of theories of consciousness currently rely heavily on theory-dependent *methodological choices* (Yaron et al., 2021).

assumption proves to be more limited in scope or strength than initially anticipated, the methodology might still have advantages compared to existing options to search for NCCs.

The implication that intrigues me most, however, is the possibility that structural approaches may introduce new *measures of consciousness*. A measure of consciousness, as conventionally understood, is a method to determine whether an organism is conscious, or whether a given stimulus or signal has been consciously perceived. Measures of consciousness are "consciousness detection procedures" (Michel, 2023) of sorts.

Building on the extensive previous work in both psychophysics and consciousness science, structural approaches raise the possibility to construct new and potentially more powerful measures of consciousness, which do not only focus on whether a single stimulus is experienced (a single quality of phenomenal character, that is), but on phenomenal character more comprehensively.

The potential of structuralist approaches in this regard can be nicely illustrated by considering verbal report, which is a paradigmatic (albeit often criticised) measure of consciousness. In the case of report, subjects use language to report facts about their experience. They might, for example, indicate that they experienced a red colour, or saw a face in a masked stimulus. The problem with reports is that when compared with the actual experience, they contain very little information. Which shade of red did the subject experience, precisely? How did they experience the face, and with which details? What else did they experience in addition to the reported fact? In information-theoretic terms, this problem arises because the channel capacity of verbal report and other behavioural indicators is low compared to the information content of conscious experiences.[7]

Structural approaches allow us to bypass the limited channel capacity of reports and similar measures of consciousness, because structural descriptions can *store information* about the phenomenal character of a subject. That is the case because structural descriptions represent features of a subject's phenomenal character that relate individual non-structural facts.

Given the structural information in a phenomenal space, a few bits of information collected in an experimental trial, for example by means of reports or similar measures of consciousness, can suffice to pin down the location in a structure, resulting in information about what a subject is experiencing that might go far beyond the bits of information that were collected. This is similar to how a geographic map can be used to decode rich information about a path based on a few bits of information about location. Finding one's way in the wilderness without a map or map-like tools generally is a very difficult task. Given a map, procedures like triangulation are available that only require a few bits of information, such as the angles between three landmarks in line of sight, to pin down one's position and find one's way. That is possible because maps store information about geography. Another example of this sort is quantum tomography, where a set of carefully chosen measurements together with structural information about the quantum state (specifically, the inner product and projective structure of the Hilbert space), is used to pin down the exact state among an infinite number of possibilities.

In a similar vein, phenomenal spaces might be used to decode information from carefully chosen low-channel-capacity measures of consciousness. Precisely how to do this remains an open question as of yet, and strongly depends on a thorough understanding of phenomenal structure in the first place (cf. Section 5), but it is a viable possibility.

2.3. Conceptual work

Structural approaches can also be essential, finally, in *conceptualising consciousness and its potential problems*. It is not unlikely that interesting philosophical implications arise, specifically in the context of structuralist assumptions, but what I'd like to highlight here is the importance of structural thinking in shaping our pre-theoretic problem intuitions about consciousness; those intuitions, that is, which guide both our theorising and experimental work.

Structural thinking might well turn what we previously thought about consciousness upside down. It might change how many of us think about our own research in the first place. To give two very preliminary examples, I think that structural approaches are relevant for epistemic arguments like Mary's room (Jackson, 1998, 1986), and for modal arguments like colour inversion (Shoemaker, 1982, Block, 1990).

For epistemic arguments such as Mary's room, the big question is whether one presumes that structural facts about experiences are known. If Mary propositionally knows, for example, which structure the experience of red has, and if structure is sufficient to individuate experiences, then she might be able to use her advanced neuroscience knowledge to create an embedding of the structure of red experiences within her own phenomenal space, even if she never experienced red, or any colour for that matter, before. Similarly, outside the realm of thought experiments, we might use structural facts to create experiences that approximate what it is like to be a bat. Structure might furnish an objective phenomenology (Lee, 2022).

Modal arguments, similarly, need to be rethought. The typical colour inversion thought experiment presumes fairly homogeneous colour spaces—colour spaces that possess symmetries. This presumption is critical because if a colour inversion is not a symmetry, then the difference between colour experience before and after the inversion will manifest itself both in behaviour and in the use of colour words: through similarity judgements and other expressions of structural facts. The closest approximation we have to a space of consciously experienced colour qualities is the CIELAB colour space (Schanda, 2007), a rendering of which is depicted in Figs. 1, 3, and 4, which is highly non-homogeneous and may not admit symmetries to the extent that we expect. Adding valence and other consciously experienced attributes of colour experiences might further erode any remaining symmetries. Thus, at least the usual intuitions regarding qualia inversions and other modal arguments may cease to be valid. Structural approaches might force us to reconsider intuitions that are built on these types of arguments.

[7] I am very grateful for a conversation with Lucia Melloni about the problems of reports and structural ideas to resolve these during a walk at the above-mentioned *Structuralism in Consciousness Studies* workshop. The idea sketched here came up during this walk and is Lucia's as much as, or even more than, mine.

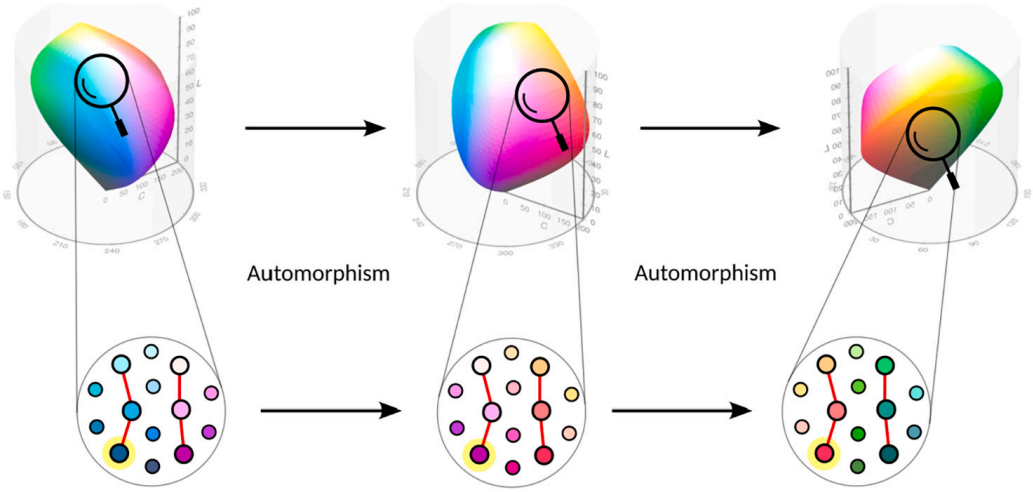


Fig. 1. What is an automorphism? This figure illustrates the concept of automorphisms. Automorphisms are somewhat analogous to rotations of a space around some axis (top row). More formally, an automorphism is a function that maps every point of a mathematical space to a different point of the same space, one-to-one, in such a way that all relations of the space are preserved: whenever two points are related before the mapping, they are also related after the mapping. This is illustrated by the bottom row, where individual points of the space are depicted by coloured dots, and relations are depicted by red lines. An automorphism maps every dot to a new dot, represented here by the change in location of the colours, in such a way that when two dots were related before the mapping (red line between two dots) the targets of the mapping are also related (red line between target dots). Automorphisms form a group because automorphisms can be inverted, and because any two automorphisms can be combined to form another automorphism, in this case one from the left-hand space all the way to the right-hand space. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article. Depiction of CIE colour space gamuts by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image is shared under the same license.)

2.4. Limitations

While structural approaches do, in my view, offer a number of benefits to the science of consciousness, it is also important to see their limitations.

A first limitation of structural approaches is that it is not clear, at present, how much of phenomenal character—how much of what it is like to experience something, that is—can be grasped by structural tools. While it is clear that much of the phenomenal structure that is usually associated with the content of consciousness can be represented structurally (much of it actually *is* structural, one might say), it is not clear whether some of the more subtle or remote facets of phenomenal character are amenable to a structural analysis. Can the experience of a self or ego be represented structurally? What about the experience of other minds? Or the pre-reflective and pre-conceptual awareness of being aware, sometimes referred to as subjective character?

A second limitation of structural approaches relates to measurability. Even if a facet of phenomenal character is amenable to structural tools, it might still be difficult, costly, or even impossible to measure. It might take years to construct a full quality space of a single modality. Is this actually feasible in experimental practise for anything but the most salient structures of phenomenal character?

A third limitation is the question of whether structural approaches can actually get any closer to modelling what is sometimes described as an intrinsic nature of qualia or qualities. Do structural approaches have any handle on modelling this? Or can they just circumscribe the structure that intrinsic properties instantiate? And to the extent that such intrinsic nature is the core of the problem of consciousness, can structural approaches get us any closer to understanding this core?

My own view of these limitations is that they define some of the key questions that structural research will have to tackle in the upcoming years. Because experiences exhibit structure, structural approaches are, by necessity, part of any research programme that targets experiences in full. To what extent they contribute to resolving the core questions at the heart of consciousness science is an open question.

3. Metaphysical premises

My first comment concerns an intuition which I have often encountered when discussing structural approaches with colleagues: that structural approaches are metaphysically presuming. Most notably, to many they seem to be tied to physicalist or reductionist metaphysics. The goal of this comment is to show that this is not the case. Structural approaches offer a new descriptive tool that can—in theory, at least—be applied independently of metaphysical assumptions, and in research programs of any metaphysical flavour. Structural approaches do not in themselves have metaphysical premises, and they do not by themselves come with a preferred metaphysical interpretation. Rather, they can be applied to and combined with the particular metaphysical ideas or presumptions that are already employed in a research program.

The major reason why structural approaches are often taken to be metaphysically presuming is that they are conflated with structuralist approaches. Structuralist approaches assume that individuals can be individuated by structure: that for every individual x, there is a unique location in a structure, a location in which only x holds. Intuitively speaking, the idea is that specification of all structural facts suffices to also specify all facts about individuals in that structure.

In the context of consciousness science, the individuals in question can be experiences, phenomenal character, qualities or qualia. The structures in question are experience spaces (spaces whose elements are experiences), phenomenal spaces, quality spaces or qualia spaces. Furthermore, there are ontological, epistemological, and methodological ways of reading a structuralist claim. In all cases, the idea is that the domain of individuals exhibits structure, and that this structure is sufficient to individuate the individuals in the relevant sense.

Structural approaches, in contrast, are not committed to a claim of individuation. An approach is structural if it applies mathematical structure. And, as I will now explain, more often than not, mathematical structure does not individuate individuals. In order to see why, we must differentiate between two readings of the term 'structure'. This will also yield a clear, formal definition of structuralism in a given consciousness-related domain.

Mathematics offers an unambiguous definition of what a structure is. A *mathematical structure* consists of two things: domains, on the one hand, and functions or relations, on the other hand. The domains of a structure are the sets on which the structure is built. They comprise the points, or elements, in a space, the individuals in a structuralist sense. In the case of a metric space, for example, there are two domains: the set of points of the metric space and the real numbers that constitute the 'distances' between points. In the case of a partial order, there is just one domain: the domain of elements that are to be ordered. The second ingredient of a mathematical structure are functions and/or relations. Functions map some of the domains to other domains. In the case of a metric structure, for example, there is a metric function that maps two points to a real number. Relations relate points to each other. In the case of a partial order, for example, there is a binary relation on the set of points. This relation specifies ordered pairs of points, usually written as $p_1 \leq p_2$.

When the term 'structure' is used in natural science, it usually follows this mathematical definition. For example, if we talk about the structure of space-time, we mean the mathematical structure that describes space-time, called a Pseudo-Riemannian manifold. If we talk about the structure of a neural network, we mean the mathematical structure of the directed graph that specifies the connectivity of the network: the mesh of nodes and edges, where each node represents a neuron or neuronal assembly, and where each directed edge specifies a neural pathway between neurons or assemblies.

When we use the term 'structure' in the context of structuralist ideas, however, it only refers to the second ingredient of a mathematical structure: the functions and relations that a mathematical structure contains. These functions or relations are what individuates the individuals—the elements of a domain—in a structuralist sense.

While customary in the context of structuralist assumptions, this use of the term 'structure' to designate only relations and functions is problematic. That is the case because we cannot actually specify relations or functions without specifying the points or elements that the relations or functions operate on. The symbol '$\leq$', for example, can be used to indicate a type of structure, a partial order in this case, but it cannot define or specify a structure. Any concrete definition or specification of a partial order needs to make use of, or refer to, the points that the relation relates. It needs to make use of some set of points—some domain in the mathematical sense of the term. Strictly speaking, it does not make sense to use the term 'structure' to refer *only* to the functions or relations. I will refer to structure in the structuralist sense—that is to the functions and relations that are part of structure in the proper sense of the term—as *structure in the narrow sense of the term*.

The structuralist idea that relations or functions determine all individuals still makes sense, of course, independently of terminological issues. And it can be expressed in a neat formal requirement, making use of the notion of an automorphism, cf. Fig. 1. An automorphism is a one-to-one mapping from the domains of the structure to themselves which preserves the functions or relations. That is, it preserves structure in the narrow sense of the term. For every point of the structure, an automorphism specifies a point as its target in such a way that the functions and relations of the structure do not change when going from the source to the target: whenever some points satisfy a relation before the mapping, they also satisfy the relation after the mapping, and equally so for functions.

Automorphisms may or may not exist. The identity mapping (not changing anything) is always an automorphism, but depending on how rich or complex the structure in the narrow sense of the term is, there might not be other automorphisms. In particular, if it is indeed the case that every point x of a structure satisfies a unique location of structure in the narrow sense of the term, then there is no automorphism other than the identity. One cannot exchange any two points without changing structure in the narrow sense of the term. In this case, one says that the *automorphism group is trivial*.[8] Vice versa, if the automorphism group of a structure is trivial, then every point must have a unique location.[9]

As structuralism (in the context of consciousness) is the assumption that every point x of a structure (in the general sense of the term) satisfies a unique location of the structure in the narrow sense of the term, structuralism is equivalent to the condition that the automorphism group of the relevant structure (in the general sense of the term) is trivial. This constitutes a nice formal characterisation of structuralism in consciousness science:

(STR) Structuralism about a domain is true iff the automorphism group of that domain is trivial.

[8] It is 'trivial' because that's the simplest possible case, and the set of automorphisms is a group because automorphisms can be combined and inverted as required by the axioms of a group in mathematics.

[9] For every point to have a unique location in a structure is for there not to exist a permutation or other mapping of the domains of that structure to themselves that leaves the structure in the narrow sense of the term invariant.

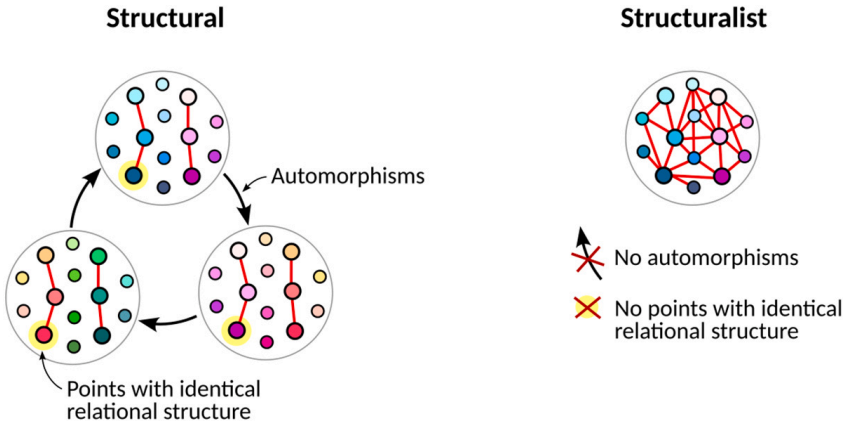


Fig. 2. Structural vs. structuralist approaches. Structural approaches make use of mathematical spaces or mathematical structures to represent or describe conscious experiences. These spaces and structures may, and in general do, admit for automorphisms (cf. Fig. 1). This implies that there are points in the space which have the exact same relational structure. Structuralist approaches, on the other hand, assume that all points of the space can be individuated by their relational structure, meaning that no two points have the same relational structure. This can only hold true if the space does not admit automorphisms, other than the identity mapping that is always an automorphism.

Here, the domain could comprise individual experiences, phenomenal characters, qualities or qualia, depending on which type of structuralism is under consideration.[10]

The crucial point of this section is that mathematical structures can, but need not, obey structuralist assumptions; they may or may not have a trivial automorphism group. A theory or experiment can be *structural*, in the sense that it makes use of mathematical spaces or structures, without necessarily being structuralist. This is illustrated by Fig. 2.

In fact, if we look at mathematical spaces in mathematics, physics, and other natural sciences, in the majority of cases, the automorphism group is *not* trivial. Simple examples of spaces with non-trivial automorphism groups are the Euclidean spaces $\mathbb{R}^2$, $\mathbb{R}^3$ and $\mathbb{R}^n$ for any $n \geq 2$, and many metric spaces, Riemannian manifolds, Hilbert spaces, or graphs.

Therefore, not only is there a difference between structural and structuralist approaches, but it is in fact quite common that the former applies while the latter does not. Structures in the general sense of the term may, but often do not, amount to structures in the narrow sense of the term. This has three consequences for research in a structuralist turn.

Consequence 1. Structural vs. structuralist agendas

Much like the two senses of the term ‘structure’ at issue here are often conflated, so are structural and structuralist agendas. Both are subsumed under the general heading of ‘structuralism’, for example. A first consequence of the above is that there is a difference between structural and structuralist agendas, and it is important to be clear about which agenda one is pursuing when engaging in structuralist research.

If one is using mathematical tools and methods, for example, to help place “structural phenomenal properties at the core of the science of consciousness” (Chalmers, 2023), as required by a very attractive position called weak methodological structuralism that has recently been put forward by David Chalmers, then one is engaging in a structural agenda: an agenda which makes use of mathematical spaces and mathematical structure but which is not committed to a structuralist claim. Put differently, structural tools like mathematical spaces can also be employed if one rejects the idea that structure (understood in the narrow sense of the term) is all that matters. They are free of explanatory and epistemic charge.

Consequence 2. Metaphysics of the mind

Many structuralist approaches are not metaphysically neutral. They imply that certain properties which some consider crucial with respect to consciousness do not exist, or are not knowable. For example, if ontic phenomenal structuralism is true, then there are no intrinsic phenomenal properties, and no genuinely private properties. Ontic structural realism implies that there are no qualia as conventionally understood (Dennett, 1988). If epistemic phenomenal structuralism is true, then one cannot know (either scientifically, or by introspection) of intrinsic or private properties, all we know about in regard to conscious experiences derives from structural properties.

Structural approaches are not tied to these assumptions. They are perfectly compatible with the existence of intrinsic or private properties. As far as the mathematics is concerned, if private or intrinsic properties exist (or if there are properties which are not accessible to structural cognitive processing), this simply means that the automorphism group of the structure is not trivial. There are points that cannot be individuated by structure alone.

[10] The term ‘domain’ also has two meanings: The meaning of domain in the sense of mathematical structure as introduced above, and the meaning of domain as a group of related items in general language. Both meanings apply here if it is clear what the structure of a domain is.

To give a very simple example, consider the case where there is no structure in the narrow sense of the term at all, i.e. the case where there are no relations or functions between qualities or qualia at all. This case can be described in terms of mathematics: the qualities or qualia simply form a set. A set is a mathematical structure according to the definition of mathematical structure in mathematical logic. It is the simplest case of a mathematical structure, but an important one. So while this case is opposed to the ideals of structuralist thinking, it is a simple but perfectly fine example of a structural approach.

What is more, structural approaches might actually help to address intrinsic, private or ineffable properties in scientific contexts. My first paper on consciousness, (Kleiner, 2020b), is devoted precisely to this issue. In a nutshell, I show that mathematical tools can be used to formulate theories of consciousness that address these properties even if they are, in an intersubjective sense, non-collatable. Because of these mathematical tools, mathematical approaches allow us to go further than non-mathematical approaches can go. Ultimately, this works because "[m]athematics translates concepts into formalisms and applies those formalisms to derive insights that are usually not amenable to a less formal analysis" (Jost, 2015).

Consequence 3. Metaphysics beyond the mind

The third consequence, finally, concerns the conviction mentioned at the beginning of this section that structural approaches seem to many to be tied to physicalist or reductionist metaphysics.

The intuition that motivates this conviction arguably derives from the equivocation of structural and structuralist assumptions, together with the idea that science can only explain relations. If structural assumptions would indeed imply that "[t]here is nothing to specifying what something is over and above stating its location in a structure" (Fink et al., 2021), and the physical sciences could only explain structure, then it would indeed be the case that structural approaches would render consciousness amenable to scientific and arguably physicalist explanation. What is more, when ontology is concerned, structuralist assumptions imply that none of the prototypical non-physicalist properties of consciousness exist (cf. Consequence 2). This, too, intuitively speaks in favour of a physicalist and reductionist research programme.

While it is clear that these intuitions do not have the force of a logical argument, it seems fair to say that *structuralist* assumptions are well aligned with physicalist metaphysics, and in the form of one of its most promising incarnations, neuro-phenomenal structuralism (Fink et al., 2021, Lyre, 2022), might even "open an attractive door for reductionism" (Fink et al., 2021).

The problem with the conviction mentioned above is that structural approaches are not necessarily structuralist approaches. The majority of mathematical spaces that are used in the sciences have a non-trivial automorphism group and therefore do not satisfy the defining criterion of a structuralist approach in the context of consciousness science (cf. Fig. 2). In other words, one can choose to apply mathematical tools and methods to describe consciousness without committing to structural assumptions and a fortiori without committing to physicalist or reductionist metaphysics. Structural approaches can be used and might be beneficial in any type of metaphysical programme, from reductive physicalism to property dualism or idealism.

In fact, there are a number of structuralist approaches which target non-physicalist metaphysics already, on the level of toy models. Atmanspacher (2020), for example, uses mathematical tools to outline how the neutral domain in a Pauli-Jung style dual aspect monism might relate to the mental and physical aspects. Other proposals, for example (Signorelli & Wang & Coecke, 2021, Signorelli & Wang & Khan, 2021), use a category theory-based graphical calculus to expand ideas from Buddhist philosophy.

In making these points, I am not to arguing for a non-physicalist research programme. My point is that structural approaches are not tied to physicalist or reductionist assumptions. Mathematical spaces and mathematical structures provide descriptive tools that can be applied to any choice of metaphysical assumptions, and in research programmes of any metaphysical flavour. Structural approaches do not have metaphysical premises, and they do not come with a preferred metaphysical interpretation.

4. Isomorphisms and structure-preserving mappings

The core question which drives the scientific study of consciousness is the question of how conscious experiences and 'the physical' relate. A ubiquitous mathematical object in the context of mathematical structures is that of an *isomorphism*, illustrated in Fig. 3 and explained in detail below. Due to its ubiquity, when introducing structure to the phenomenal domain, many feel that it is natural to assume that this structure is related to physical structure by an isomorphism or structure-preserving mapping. My goal here is to show that this assumption is not in fact justified. We either need to search for a rigorous justification, or if there is none, proceed in different ways.

Intuitively speaking, an isomorphism expresses a relation between two structures. Precisely speaking, it is a bijective mapping *between the domains* of two structures that preserves the relations or functions of these structures. That is, it is a map from the elements or points of one structure to the elements or points of another structure. A map is bijective if it is one-to-one and onto, meaning that every element in the target space gets mapped to by exactly one element in the source space.

In practice, because the physical has a much larger domain and much richer structure than the phenomenal, when the concept of an isomorphism is applied in consciousness science, what is actually meant is an *isomorphism onto the image*. This means that there is an isomorphism from the phenomenal domain to a substructure of the physical domain. Often, homomorphisms are used as well. They are defined exactly like isomorphisms, except that they do not have to be one-to-one or onto, so that some elements in the target space might not get mapped to, and/or several elements in the source space might map to the same element in the target

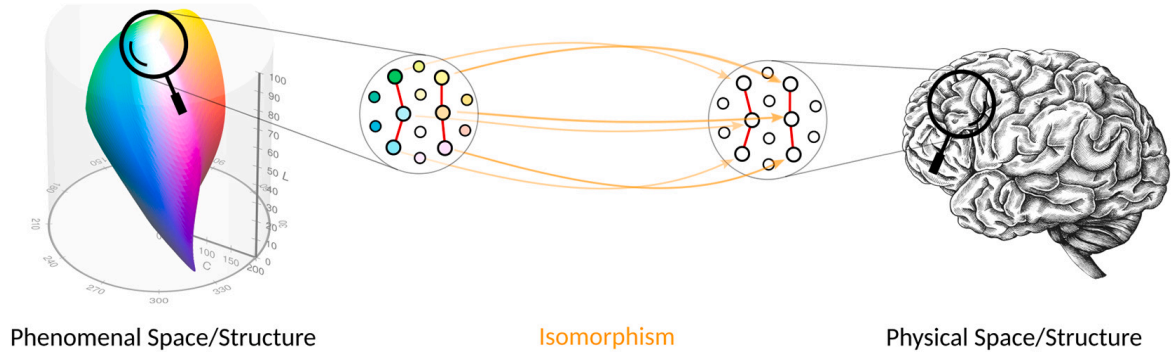


Fig. 3. What is an isomorphism? This figure illustrates the concept of isomorphisms as applied in consciousness science to link a phenomenal space or structure (left) with a physical space or structure (right). By definition, isomorphisms operate on the level of points. An isomorphism maps every point of the phenomenal space to a point in the physical space. It does so in such a way that the relations between points (indicated here by red lines) are preserved, meaning that any two points which are related on the left are related in the exact same way on the right. The mapping also needs to be invertible. An isomorphism presupposes that structures on both sides of the mapping are given. It does not define, or pick out, the structure in its target domain, which is why it is not a suitable mathematical object to explain, predict, or define phenomenal structure in terms of physical structure. (Depiction of CIE colour space gamut by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image, excluding the drawing of the brain, is shared under the same license. Drawing of the human brain from Freepik.)

space. Strictly speaking, though, the mathematical concept of homomorphisms is not appropriate either,[11] but to avoid unnecessary technical details, I will admit them too. I will use the term *structure-preserving mapping* to denote homomorphisms or isomorphisms with the understanding that the domains and structures of the source and target have been adapted appropriately to avoid the technical problems. As far as intuition is concerned, my comments are easiest understood when thinking about an isomorphism onto the image.

The assumption under discussion then is:

(ISO) The physical and the phenomenal are related by a structure-preserving mapping from the phenomenal domain to the physical domain.[12]

This assumption is a very consequential assumption. It promises, for example, a new methodology for measuring Neural Correlates of Consciousness (NCCs). To date, NCC research has to make use of intricate measures of consciousness (Irvine, 2013), to distinguish between trials where the subject perceives a stimulus consciously from trials where it doesn't. If (ISO) is true, a whole new avenue for investigating NCCs is available: to search, among neural structures in the brain, for structures that are homomorphic to or identical with the structures of the phenomenal domain. This search could, in principle, be carried out independently of any measure of consciousness, and might give a unique result, so that potentially at least there is a methodology where one "[does] not have to worry whether subjects 'really' had a phenomenal experience of a stimulus" (Kob, 2023).

The existence of a structure-preserving mapping between the phenomenal and physical domain also has important consequences for theories of consciousness: it implies that a large class of theories of consciousness is false, namely all those which do not take the form of a homomorphism. A good example of this is Integrated Information Theory (IIT) (Oizumi et al., 2014, Albantakis et al., 2023). It is sometimes assumed that IIT is structure-preserving or even an isomorphism, but according to IIT's mathematical formulation, this is not the case. The mathematics of IIT come with two clear 'slots' for the physical and phenomenal domain. One of the slots is the input to the theory's algorithm. It requires a physical description of a system, for example in terms of neurons. The other slot is the output of the theory's algorithm. For every system and physical state of this system, this output is a mathematical structure called 'Maximally Irreducible Conceptual Structure' in IIT 3.0, and 'Φ-structure' in IIT 4.0. This structure "is identical to [the system's] experience" (Oizumi et al., 2014). The mathematical algorithm of the theory specifies a mapping between those two

[11] The concept of homomorphism as used in mathematics presumes that two structures have the same *signature*, meaning that both structures need to have the same type of functions or relations: the same number of functions or relations of the same arity, that is. Because the physical has much more structure than the phenomenal (think about the rich structure of electrodynamics in the case of neurons, say), the concept of homomorphism is too strong to express the underlying idea. One could attempt to define a *partial homomorphism* as a homomorphism that respects some, but not all, structures of the target domain. But for questions other than multiple realizability, the 'isomorphism onto the image' conception seems to be closer to the underlying intuition. The same applies if one reverses the direction of the homomorphism, cf. Footnote 12.

[12] In addition to the problem mentioned in Footnote 11, there is also the question of which direction a homomorphism should take. Should it go from the physical domain to the phenomenal domain, as in Fink et al. (2021), or vice versa? Because it is unlikely that all elements of the physical domain are mapped to the phenomenal domain (there are neural mechanisms which are not relevant for conscious experiences, for example), and because a map in the sense of mathematics requires a specification of a target element for every element of the source domain, it seems more natural to me to choose the phenomenal-to-physical direction. Choosing the physical-to-phenomenal direction would require one to introduce yet another sense of partiality, that of a partial function, which is only defined on some of its elements. The problem with this is that a homomorphism which is partial in both this sense and the sense of Footnote 11 always exists, so that the statement (ISO) is vacuous. This is not the case for an isomorphism onto the image in the phenomenal-to-physical direction, because of the need to specify a target element in the physical for every source element in the phenomenal in such a way that the image has the same structure as the phenomenal. This is why I think isomorphisms onto the image in the phenomenal-to-physical direction are the right tool (and the right intuition) to work with, though my comments below do not turn on this choice.

slots which is not a homomorphism. Therefore, the theory does not specify a homomorphism between the physical and phenomenal domains. And consequently, if (ISO) is true then IIT must be wrong.[13]

4.1. Are isomorphisms justified?

The above shows that (ISO) is indeed a very consequential assumption. This would be good news if (ISO) were also a justified assumption. However, as I will argue here, this is not the case. While isomorphisms and homomorphisms are natural in mathematics, they appear not to be the right sort of object to achieve the goals of consciousness science in investigating how the phenomenal and the physical relate. For the purpose of this discussion, I will assume that these goals are "to *explain*, *predict*, [or] *control* the phenomenological properties of conscious experience" (my emphasis) in terms of physical properties, following Anil Seth's *Real Problem of Consciousness* (Seth, 2021), with the understanding that phenomenal structure is an integral part of phenomenal character, and that structural properties are properties too.

My comments are tied directly to what an isomorphism or homomorphism is. As explained above, isomorphisms and homomorphisms are mappings between the domains of two structures (between the *points* or *elements* of these structures, that is) which satisfy certain conditions. The conditions enforce that the mappings are compatible with the structures on both ends. This has two important consequences for the question at hand.

The first consequence is that a homomorphism *presupposes* that the structures on both ends of the mapping are given. If only one of the two structures is given, or none even, then (ISO) becomes an empty statement. This is because *any* mapping of the form $f : E \to P$, where P denotes the physical domain and E denotes the experiential domain, can be turned into a homomorphism if at most one domain comes with structure. One can simply define the structure on the other domain so that the mapping is a homomorphism. Assuming that there is a homomorphism without presupposing that structures on both ends of the mapping are given amounts to not assuming anything at all.

But if a homomorphism presupposes structures on both ends, it doesn't explain, predict, or allow to control these structures. Homomorphisms fall short of explaining, predicting, or controlling those phenomenal properties they were introduced to cope with.

Second, and more importantly in my opinion, homomorphisms do not have the right mathematical form to *pick out* which structure there is. That is the case because they are maps from domains to domains. They do not actually map from structures to structures, as is sometimes thought. They only map points in one domain to points in another domain in such a way that the mapping between the points *preserves* or *respects* the structure on both ends. This speaks against an explanatory or predictive function as well, as I shall now explain.

Let us first consider the case of explanation. Do homomorphisms, or other structure-preserving mappings, *explain* phenomenal structure in terms of physical structure? There are various notions of explanation that are available in science, ranging from the early deductive-nomological and inductive-statistical ideas studied by Carl Hempel (Hempel & Oppenheim, 1948, Hempel, 1962) to more modern understandings of explanation in the form of causal-mechanical models (Salmon, 1984), unificationist models (Friedman, 1974, Kitcher, 1989), contrastive explanation (Van Fraassen, 1980) or interventionalist models (Woodward & Hitchcock, 2003, Hitchcock & Woodward, 2003).

It is clear that homomorphisms do not fit the original Hempel models of explanation because they do not derive phenomenal structure in any meaningful sense from a general law and initial conditions. What is crucial though is that they also don't sit well with the other models of explanation. This is the case because, in one form or another, these models all require 'what if things had been different' information. In the causal-mechanical model of explanation, 'what if things had been different' information is required to test the robustness of a purported causal mechanism. In unificationist models it matters for questions of breadth of a unifying explanation. In contrastive explanations it is central to deal with alternative scenarios that would have occurred under different conditions. And in interventionist models, it is required to explicate how an intervention changes the explanadum variable.

Homomorphisms do not pick out structure on the physical or phenomenal side, they only relate points of the domains in a structure-preserving way. Therefore, they do not provide 'what if things had been different' information about phenomenal structure. But 'what if things had been different' information is required by the above-mentioned models of explanations. Therefore, homomorphisms do not constitute an explanation of phenomenal structure according to these models.

Because homomorphisms don't pick out phenomenal structure, they do not offer alternatives to how phenomenal structure could be if things had been different. For this reason, they also do not predict phenomenal structure. Prediction, too, requires mathematical tools that pick out the right structure among a class of possible structures.

A helpful way to think about the problems of explanation and prediction is to think about what would *define* phenomenal structure in terms of neural structure, or physical structure more generally. Consider, as an analogy, computer games. Computer games employ

[13] The only way to enforce viewing IIT as an isomorphism is by claiming that the output of IIT's algorithm is itself a physical structure, which then happens to be related by an isomorphism to the phenomenal domain. Given the interpretation of the mathematical structure outputted by IIT as "identical to [the system's] experience" (Oizumi et al., 2014), it is hard to see how such interpretation can plausibly be made. The mathematical quantities outputted by IIT do not appear anywhere else in the physical sciences, and are conceptually and mathematically rather removed from physical theories. Such a claim also violates the implicit presupposition in (ISO) that there are more or less well-defined structures on both the phenomenal and physical sides. If there were no constraints on which structure to consider, then (ISO) would be a vacuous statement. Any mapping of the form $f : P \to E$, where P denotes physical structure and E denotes phenomenal structure, can be turned into an homomorphism between the physical and the phenomenal if E is taken to be a physical structure as well. As a rule of thumb, if a structure is actively defined by a theory of consciousness, rather than just adapted from some other part of science, it should probably not count as physical structure in the sense required by (ISO).

mathematical structure to model rich and detailed visual imagery. Yet the mathematical models are defined mostly in terms of objects in the sense of object-oriented programming. There is nothing in the actual code of the game which resembles the structure of the visual scene; rather, the code defines how the structure should be rendered, and it does so in terms of objects and properties. The visual structure created by the game is not homomorphic to the code that runs in order to create the scenes, yet it is defined by the code. This example illustrates that homomorphisms are not the kind of thing one would expect when defining structure.

What these points illustrate, in my view, is that homomorphisms and structure-preserving mappings more generally are not the right sort of object to define, explain, predict, or control phenomenal structure. They might be natural in the context of mathematical questions, but they are not natural for the purposes of consciousness science.

Consequently, (ISO) is not in fact a natural or justified assumption. We either need to search for a rigorous justification, or if there is none, proceed in different ways. Because (ISO) is so consequential for theoretical and experimental work, using (ISO) without proper justification, or in the hope that a justification will eventually be found, is not a viable option.

This comment also applies to mathematical objects known under different names, if these objects are in fact homomorphisms. Important examples thereof are diffeomorphisms, which are maps between smooth geometric shapes called manifolds. Diffeomorphisms are homomorphisms between the mathematical structures that define smooth manifolds. And much like the simpler cases discussed above, they map points of one manifold to points on another manifold in a way that respects the mathematical structure on both sides of the map. They do not explain or define the structure.

4.2. What, if not isomorphisms?

If isomorphisms and homomorphisms are not the sort of thing that explains, predicts, or defines phenomenal structure, what is? Which mathematical objects should we use to relate the physical and the phenomenal in a structuralist turn?

My view is that there is no general mathematical principle that we can commit to. Rather, much like theories of consciousness in the pre-structural area were built one-by-one, we have to build structural theories one-by-one, working with different ideas, concepts, motivations and metaphysics in each case. The challenge of finding the right mathematics to explicate these ideas, concepts and motivations in a structural context is not something we can bypass by choosing one mathematical tool that fits them all. This is not technically possible, nor is it desirable. The difference between ideas, concepts and metaphysical underpinnings in a structural context is precisely in the mathematics that relate the physical to phenomenal structure. We cannot waive the problem of finding the right mathematics without also waiving the possibility of choosing different metaphysical or conceptual ideas.

5. Which phenomenal structure?

My final comment concerns the question of which structure to consider when embarking on structural research. That is the question of what phenomenal structure *is* and how we find it. This question is important because conscious experience does not "come with" mathematical structure in any direct sense. There is nothing in what it is like to experience something that is per se mathematically structured, other than if one explicitly experiences something mathematical.[14]

Rather, mathematical spaces and mathematical structures are *tools* or *languages* we can use to describe (or model) phenomenal character, much like English or any other language can be used to describe phenomenal character. And just as we need definitions or conventions to apply English language terms, we need definitions or conventions to apply mathematical terms. These might not be as simple as in the case of English, but still they flesh out the conditions under which one is, and under which one is not, justified in making a structural and mathematical claim.

Because mathematics is a different type of language from English, the definitions or conventions to apply structural terminology are of a different type too. They constitute *methodologies*, meaning they are collections of methods, procedures or rules, that can and need to be used to assess mathematical claims.

And because phenomenal character does not "come with" mathematical structure in any direct sense, any claim about a structural fact, and any application of structural ideas, is always *relative* to a specific understanding of what phenomenal structure is, and a fortiori, relative to the methodology that defines this particular understanding. It is not meaningful to claim that experiences have a certain structure. Much like a claim about whether experiences have qualia depends on what exactly one takes the term qualia to denote, the claim that experiences have a certain structure depends on what one takes phenomenal structure to denote (Fig. 4). When working with or thinking about phenomenal structure, we need to be clear about which methodology we presume. Otherwise, we are prone to making errors. This is the first major point I would like to make in this comment.[15]

[14] We do experience mathematical structures if we know and recognize them, for example in the case of geometrical shapes, or if we actually work with mathematical structures. But we do not experience non-mathematical experiences as mathematically structured. We do not, for example, experience colours as constituting a metric space or having a partial order.

[15] Therefore, working with mathematical structure in consciousness science is different from working with mathematical structure physics or other natural sciences. In physics and other natural science, we do not have direct access to the phenomena we are studying. In a certain sense, for structural claims in physics, *anything goes*, as long as the relevant notion of measurement for that structure reproduces what is observed. This is why there are hugely different proposals about the structure of spacetime, for example, ranging from quantized spacetime (Rovelli, 2004) and emergent spacetime (Koch & Murugan, 2012) to proposals that depart completely from what we intuitively think spacetime should be (Finster & Kleiner, 2015). As long as limiting processes exist that relate these proposals to previous models, in this case the notion of spacetime of General Relativity, all those proposals are viable options. This is not the case for consciousness, because consciousness has a different epistemic context. For example, it exhibits what is sometimes called *epsitemic asymmetry*: there are "two fundamentally different methodological approaches that

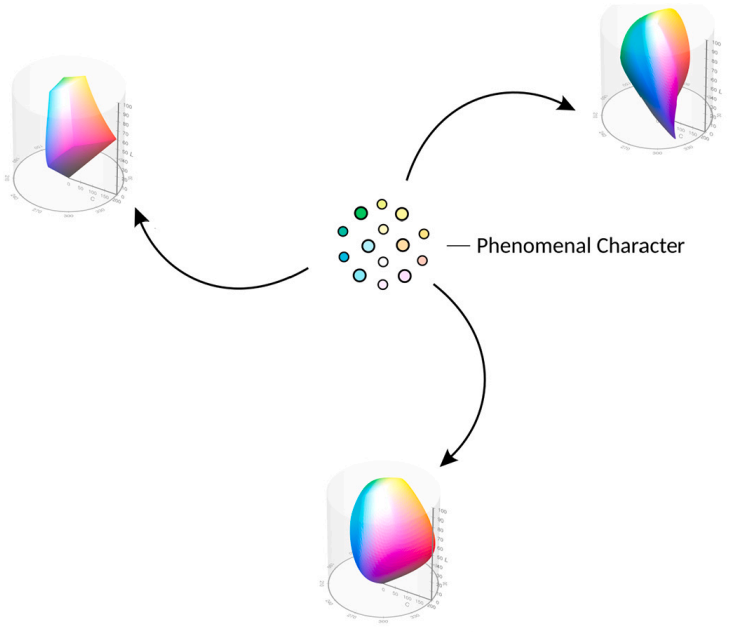


Fig. 4. Different definitions imply different spaces. Mathematical spaces and mathematical structures are tools to describe or represent phenomenal character, much like technical language terms are too. Different definitions or conventions of how to use mathematical terms to describe or represent phenomenal character—different conventions of what terms like 'mathematical structure of conscious experience' or 'phenomenal space' mean—lead to different structural representations of the same set of experiences, here illustrated by three different CIE colour spaces. Black arrows indicate different definitions or conventions, which imply different methodologies for constructing phenomenal spaces in the lab. Different geometrical shapes indicate different types of spaces that result from applying these methodologies. Much like technical language terms might differ in scope, quality, adequacy, and presuppositions, definitions or conventions regarding mathematical structures differ in scope, quality, adequacy and presuppositions. (Depiction of CIE colour space gamuts by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image is shared under the same license.)

5.1. *What is phenomenal structure, and how do we find it?*

There are three important landmarks that have influenced the way in which we use mathematical structures to describe conscious experiences today: quality spaces as introduced by Austen Clark (Clark, 1993), quality spaces as introduced by David Rosenthal (Rosenthal, 1991, 2010) and Q-spaces as introduced in IIT 2.0 (Tononi, 2008). While these methodologies have served an important function in enabling structural research, it is also important to be clear about their shortcomings in moving forward.

As far as IIT is concerned, the obvious shortcoming is that the theory does not provide a phenomenal interpretation of the structure it proposes, other than the claim that the structure "is identical to [the system's] experience" (Oizumi et al., 2014). This gives rise to what David Chalmers has called the *Rosetta Stone Problem* (Chalmers, 2023): the problem of how to translate the mathematical structure that IIT proposes into phenomenological terms. IIT does not actually specify a methodology that clarifies how to interpret and test their proposed structure in phenomenal terms.

The proposals by Clark and Rosenthal do specify methodologies. The major shortcoming of these methodologies, on my view, is that they conflate three sources of mathematical structure:

1. **Mathematical Convenience.** Some of the structure is introduced simply for mathematical convenience.
2. **Laboratory Operations.** Some of the mathematical structure refers to, or depends on, laboratory operations.
3. **Conscious Experience.** Only part of the mathematical structure actually pertains to conscious experiences or phenomenal character.

5.2. *Clark's quality spaces*

Quality spaces as introduced by Austen Clark (Clark, 1993) are based on the following methodology. To construct the quality space for an individual subject,[16] one fixes a class of stimuli $\mathcal{S}$ that can be presented to the subject, and defines two tasks that the subject can complete in response to the presentation of one or more stimuli. The first task probes whether the subject is able to

enable us to gather knowledge about consciousness: we can approach it from within and from without; from the first-person perspective and from the third-person perspective. Consciousness seems to distinguish itself by the privileged access that its bearer has to it" (Metzinger, 1995). In other words, in addition to the usual scientific way of accessing and modelling a phenomenon there is a second way of accessing the phenomenon (described in terms of the first person perspective *metaphor* above). Due to this different epistemic context, using mathematical structure to describe a phenomenon is different in the case of consciousness, and more constrained, than in the case of physics.

[16] Clark mostly has humans in mind, but does consider the case of animals briefly in Clark (1993). Nothing hinges on humans in the methodology he proposes.

discriminate the experience elicited by two different stimuli consciously. The second task probes whether the subject experiences a stimulus to be more similar to a reference stimulus than another stimulus. This is called *relative similarity*.[17]

The discrimination task is used to define a *global indiscriminability* relation on the class of stimuli $\mathcal{S}$.[18] While discriminability does not constitute an equivalence relation, global indiscriminability does. This equivalence relation partitions the set of stimuli. Each set in this partition contains stimuli which are globally indiscriminable from each other, and defines a *quality* in Clark's proposal. The collection of the sets in this partition (the space of equivalence classes of $\mathcal{S}$, in mathematical terms) defines the domain of the quality space that is being constructed.

The relative similarity task is used to define a graph, in the mathematical sense of the term, between the qualities: a set of nodes, and edges that link some of the nodes. Working with stimuli that represent the different qualities, one first collects relative similarity data. This is data about whether a quality q_1 is more similar to a reference quality q_0 than another quality q_2. One might find that the pair (q_1, q_0) is more similar to each other than the pair (q_2, q_0), say. Having collected this data for all qualities in the set, one then represents them as a graph. Every quality one has previously constructed is a node of the graph, and every pair (q_i, q_j) about which one has relative similarity data is an edge of the graph between the nodes that represent the qualities. The important part then is that the edges get labelled by numbers, and these numbers must be chosen in such a way that the relative similarity judgements that have been collected are represented truthfully by the ordering of the numbers. The label of the edge (q_1, q_0) above, for example, must be a lower number than the label of the edge (q_2, q_0) if the former pair is more similar to each other than the latter pair. The result of this procedure is a labelled graph, where the nodes represent qualities, edges indicate pairs for which similarity data is available, and labels on the edges represent relative similarity. In mathematical terms, this is called a POSET-labelled graph, where a POSET is a partially ordered set. The partial order is the phenomenal structure of the relative similarity experiences.

Up to this point all the mathematical structure is still grounded in conscious experience, to a large extent. The data to carry out the constructions is based on tasks that might utilize reports or behavioural measures, but these measures depend on what is experienced.

The next step in Clark's methodology consists of introducing a metric, a tool to measure distances in terms of continuous numbers, and in fact an Euclidean space that has a uniform, homogeneous metric. To this end, it makes use of a procedure known as 'multidimensional scaling' (Beals et al., 1968). In Clark's case, it consists of finding an *embedding* of the graph into an Euclidean metric space in such a way that the distance between the nodes of the graph—which are mapped to points in the metric space—reproduce the ordering of relative similarity that the labels of the graph encode.

From the perspective of phenomenal character, this step is unwarranted. Not only is the metric introduced without any reference to experience, but this step also leads to the introduction of many more points besides the original qualities that were carefully constructed making use of global indiscriminability. Technically speaking, it leads to an infinity of additional points, all of which feature in the metric function of the space, and none of which is any different from the points that were carefully constructed based on tasks and stimuli.

The only justification I can think of why one would make use of this last step, as compared to just working with the POSET-labelled graph, is mathematical convenience. A POSET-labelled graph might just be too unfamiliar a mathematical object. Or maybe the reason is that it cannot easily be further analysed on a computer in familiar ways. These justifications are in fact made explicit in introductory texts on psychophysics. Luce and Suppes, for example, speak of representational measurement, of which multidimensional scaling is an example, as "an attempt to understand the nature of empirical observations (...) in terms of *familiar* mathematical structures" (Luce & Suppes, 2004, p. 1) (my emphasis), and add that "the use of such empirical structures in psychology is widespread because they come close to the way data are organised for subsequent statistical analysis" (Luce & Suppes, 2004, p. 2). Be that as it may, the last step that introduces the metric function fails to be grounded in conscious experience. It is an example of 1. above.

5.3. Rosenthal's quality spaces

The construction of quality spaces as defined by David Rosenthal is based on a class of stimuli as well. But in this case, one only needs a discrimination task, as well as means to *vary* the stimuli.

The main step in Rosenthal's methodology is to construct *Just Noticeable Differences* (JNDs) from variations of the stimuli and the discrimination task. To this end, one varies a stimulus in some direction until the subject notices the difference between the stimulus and the variation. The class of stimuli which one can reach by varying one stimulus without creating a JND gives a set or region in stimulus space, and much like in the case of Clark, the idea is that these regions constitute qualities. A metric function is introduced on the set of qualities by counting the minimal number of regions one has to pass so as to go from one quality to the other.

In this proposal too, there is a question as to the experiential source of the metric function. Because the metric function can be specified, once JNDs have been constructed, without need of additional data, it might not represent anything over and above the JNDs and their neighbourhood relations. Furthermore, while we do experience colour qualities as instantiating a relative similarity

[17] There is considerable freedom in which class of stimuli to choose and how to define and implement the tasks.

[18] Two stimuli are globally indiscriminable if and only if the following two conditions hold:

1. The two stimuli are indiscriminable from each other.
2. The two stimuli have identical indiscriminability relations to all other stimuli in $\mathcal{S}$.

structure, we do not experience qualities to be a certain number of steps apart, as a metric would require if it indeed represented a structure of conscious experience.[19] So there is a worry of the metric being due to mathematical convenience here too.

A more fundamental worry in this case concerns the *variations* of stimuli that one needs in order to construct JNDs and their neighbourhood relations in the first place. The idea of a variation—starting with one stimulus and then changing that stimulus continuously until a subject notices a difference—requires a *topology* on the stimulus space. A topology defines what it means to "draw a line without lifting a pen" on an abstract space, so to speak. It is precisely what provides the notion of continuous curves required to specify variations in Rosenthal's definition. Without a topology, a variation can jump from any point to any other point.

The problem is that different topologies give different variations. So when one actually constructs a quality space according to Rosenthal's methodology in the lab, the resulting space depends on the topology of the stimulus space that has been used. And much like there isn't just a single notion of colour space, there isn't just a single topology on colour stimuli one can use. As a result, the metric function that one constructs in an application of Rosenthal's methodology actually depends on the topology that has been chosen in the experiment, which is a laboratory operation in the sense of 2. above.

In the case of Rosenthal's methodology, there is in fact a theory that can be used to answer these and similar worries, a theory about what consciousness is, about how qualities should be understood, and about how consciousness and qualities relate. When I asked David Rosenthal about the problem regarding topology, for example, he countered by assuming that there is just one actual physical topology in reality and that this is the topology that should be used. It is not clear to me how this would work in practice, given that this topology is presumably defined by Quantum Electrodynamics (QED), and too far removed from experimental practice to be applicable; in the lab, some choice of topology will have to be made nonetheless. But theoretically speaking, the answer is fully valid. Similarly, the theory about what qualities are and how they relate to consciousness discharges the methodology from the problem that, according to the subsumed notion of discrimination in this case, discriminations could also be made unconsciously.

There is, however, no free lunch. The price to be paid for solving methodological problems by theoretical assumptions is that the methodology now depends on these assumptions and cannot be used to formulate or test other theories of consciousness. The methodological tool might be deprived of much of the impact it could otherwise have.

In my view, quality spaces are ways to describe or represent the explanandum—what is to be explained: qualitative or phenomenal character, what it is like to be—, while theories of consciousness are the explanantia—they do the explaining. This is why I have always been tempted to read Rosenthal's proposal as a general methodology that is independent from his theory. This is possible and addressing the above-mentioned problems on purely methodological grounds leads, in my view, to fruitful further developments of his construction (cf. below and Kleiner and Ludwig (2023)).

5.4. How to move forward

In the last two sections, I have analysed two proposals for methodologies that define what quality spaces are. While these proposals have served an important role in enabling structural thinking, much of the essential structure in these proposals is not actually grounded in conscious experiences, but in mathematical convenience and laboratory operations.

It is possible to go beyond individual methodologies and analyse the *type of condition* that is applied in these proposals and more recent work. That is, the type of condition that decides whether a mathematical structure is a quality space or phenomenal space—*a mathematical structure of conscious experience*, to use a general term. In a nutshell, all existing proposals I know of amount to:

(A) Conditions on the domains (sets of points) of a mathematical structure, formulated in terms of qualities, qualia, phenomenal properties or similar aspects of conscious experiences.
(B) The requirement that the mathematical axioms of the structure (such as the axiom that the metric distance between a point and itself is zero) are satisfied.

This type of condition can be shown to be insufficient to ground a thorough understanding of phenomenal structure. This is the case because (a) it is prone to admitting incompatible structures, (b) allows for *arbitrary* re-definitions of structures that still satisfy the condition, and (c) in a subtle but important sense, the condition is indifferent to structural facts of conscious experience. I do not have the space here to explain these problems in detail; they are explained and illustrated in (Kleiner & Ludwig, 2023, Section 1).

I take the problems of existing proposals, and the insufficiency of the general type of condition that is applied, to constitute a need of constructing a *new methodology* for phenomenal spaces. This methodology needs to take previous methodologies into account, but needs to amend and extend them to avoid the three insufficiency problems as well as the issues with non-conscious sources of the mathematical structure.

In Kleiner and Ludwig (2023), Tim Ludwig and I have set out to find a methodology that achieves this task. The result is illustrated in Fig. 5. The proposal shares with David Rosenthal's methodology that it rests on variations, though in our case, any transition from one conscious experience to another counts as variation, and we do not demand continuity or restrict only to variations of stimuli.

Put in terms of phenomenal properties, the core intuition of our proposal is that a mathematical structure is a mathematical structure of conscious experience—a phenomenal space, to use a simpler term—if and only if there is a phenomenal property that behaves exactly as the mathematical structure does under variations. If a variation preserves the mathematical structure (if it is an

[19] For a more careful examination of the case of a metric, cf. Kleiner and Ludwig (2023). For questions on how quality spaces *should* relate to consciousness or phenomenal character according to the underlying theory, cf. below.

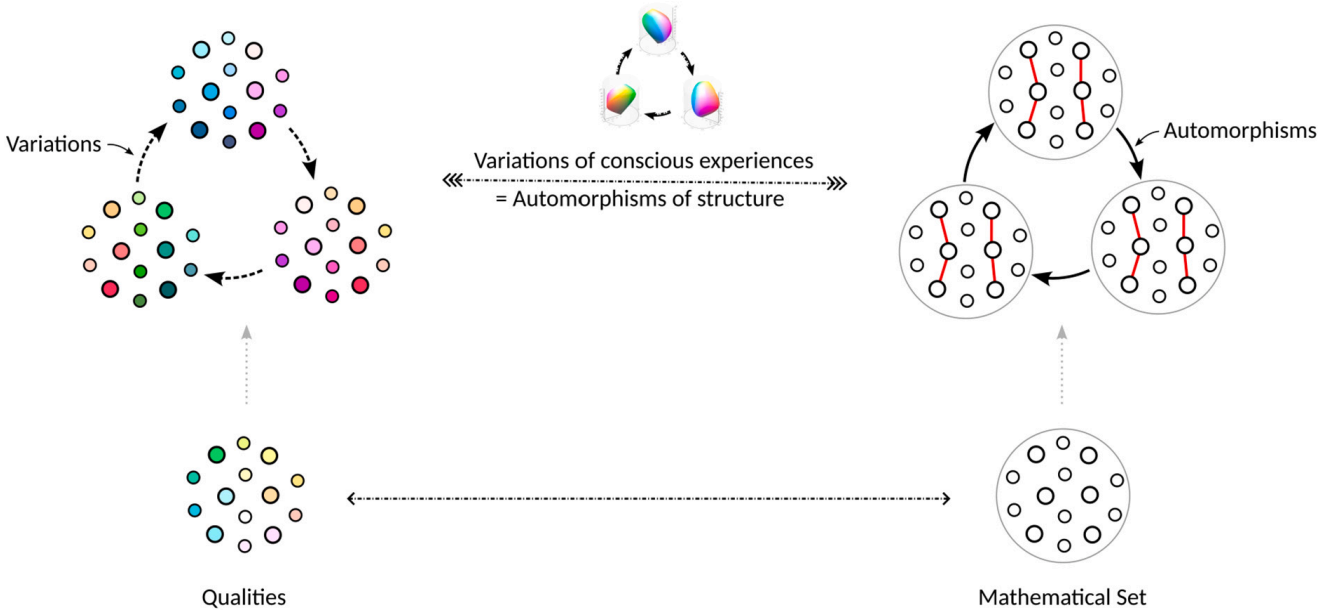


Fig. 5. How to define phenomenal spaces? This figure illustrates how to define phenomenal spaces and other mathematical structures of conscious experience. One starts out with a choice of qualities (bottom left), for example colour qualities, sometimes also called qualia or conceptualised as instantiated phenomenal properties. The qualities form a set that constitutes the points of the phenomenal space (bottom right). Every experience comprises a subset of qualities, and as experiences change from one experience to the next, the subset of qualities that is realised varies (top left). These variations can be understood as mappings from the set of qualities to itself, and therefore have the same formal structure as automorphisms (Fig. 1): mappings from the points of a space to other points of the space (top right). This allows for the following simple definition of phenomenal structure: phenomenal structure is that mathematical structure whose automorphisms are identical to the variations of the qualities as experiences change. Put differently, phenomenal structure (indicated here by red lines) is that mathematical structure which renders the statement true: the variations of (qualities of) conscious experiences are the automorphisms of the structure (top centre). For details, cf. (Kleiner & Ludwig, 2023). (Depiction of CIE colour space gamuts by Wikimedia Commons, Michael Horvath, under CC BY-SA 4.0; this image is shared under the same license.)

automorphism of the structure, in mathematical terms), then it must not change the phenomenal property. If, conversely, a variation does not preserve the mathematical structure, then it must change the phenomenal property. In a nutshell: there is something "in" conscious experience (the phenomenal property) that behaves exactly as the mathematical structure does.

6. Conclusion

Structural approaches, which make use of mathematical structure to describe or model conscious experiences, offer new and valuable avenues for studying consciousness. My aim in this paper is to provide three comments that I consider important when engaging in structural research. Each comment targets what is, in my view, a misconception or misunderstanding that I aim to clarify.

My first comment focuses on the metaphysical underpinnings of structural approaches. I show that, contrary to popular belief, structural approaches are not tied to physicalist or reductive metaphysics. Instead, they offer versatile descriptive tools that can be utilised irrespective of one's metaphysical commitments, across research programmes of any metaphysical flavour.

My second comment concerns isomorphisms and structure-preserving mappings. A number of emerging structuralist research programmes rely on assuming a structure-preserving mapping between the phenomenal and the physical domain. I argue that this assumption is unwarranted, and that isomorphisms and structure-preserving mappings are not the right mathematical object to provide explanations, predictions, or definitions of phenomenal structure. Instead, we should direct our attention to structural theories of consciousness, without expecting a single mathematical formalism to fit them all. One major experimental consequence of this is that methods such as Representational Similarity Analysis (Kriegeskorte et al., 2008), which searches for structural similarity, may not be the right approach to search for the neural correlates of phenomenal structure.

My third and final comment focuses on the question of what phenomenal structure *is*, and how we find it. Conscious experiences do not "come with" mathematical structure in any meaningful sense. Rather, mathematical spaces and mathematical structure offer a language to describe or represent conscious experiences, and just like we need definitions or conventions to apply English language terms to consciousness, we need definitions or conventions to apply structural terms. In the case of structure, the definitions and conventions take the form of methodologies that govern how to construct or use the mathematical terminology. The two major methodologies that have guided recent developments are quality spaces as introduced by Austen Clark, and quality spaces as introduced by David Rosenthal. I show that both suffer from fundamental issues, and discuss how to move forward in light of this.

CRediT authorship contribution statement

Johannes Kleiner: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

I would like to thank the organisers and participants of the 2023 *Structuralism in Consciousness Studies* workshop at the Charité Berlin for many stimulating discussions on this topic, in particular Lukas Kob, Lucia Melloni, Sascha Benjamin Fink, Holger Lyre, David Chalmers, Andrew Lee, and Wanja Wiese, as well as the participants of a recent *NYU Philosophy of Mind Discussion Group* for valuable advice, feedback and discussions of an earlier version of this manuscript. Furthermore, I would like to thank Wanja Wiese, Matthias Michel, and Moritz Nicolas Loerbroks for feedback on an earlier version of this manuscript. This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation. I would like to thank the Mathematical Institute of the University of Oxford and the NYU Center for Mind, Brain, and Consciousness for hosting me while working on this article.

Funding

This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G., Zaeemzadeh, A., Boly, M., Juel, B. E., et al. (2023). Integrated information theory (iit) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology*, *19*(10), Article e1011465.

Atmanspacher, H. (2020). The Pauli–Jung conjecture and its relatives: A formally augmented outline. *Open Philosophy*, *3*(1), 527–549.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*(2), 127.

Block, N. (1990). Inverted Earth. *Philosophical Perspectives*, *4*, 53–79.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the conscious Turing machine. *Proceedings of the National Academy of Sciences*, *119*(21), Article e2115934119.

Chalmers, D. (2023). *Phenomenal structuralism*In *Talk presented at the structuralism in consciousness studies workshop at the Charité Berlin*.

Chalmers, D. J., & McQueen, K. J. (2022). Consciousness and the collapse of the wave function. In S. Gao (Ed.), *Consciousness and quantum mechanics*. Oxford University Press.

Clark, A. (1993). *Sensory qualities*. *Clarendon library of logic and philosophy*.

Clark, A. (2000). *A theory of sentience*. Clarendon Press.

Coninx, S. (2022). A multidimensional phenomenal space for pain: Structure, primitiveness, and utility. *Phenomenology and the Cognitive Sciences*, *21*(1), 223–243.

Dennett, D. C. (1988). Quining qualia. In *Consciousness in contemporary science* (pp. 42–77).

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*.

Finster, F., & Kleiner, J. (2015). *Causal fermion systems as a candidate for a unified physical theory*. *Journal of Physics: Conference Series*, *626*, 012020. IOP Publishing.

Fortier-Davy, M., & Millière, R. (2020). The multi-dimensional approach to drug-induced states: A commentary on Bayne and Carter's "dimensions of consciousness and the psychedelic state". *Neuroscience of Consciousness*, *2020*(1), Article niaa004.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5–19.

Gert, J. (2017). Quality spaces: Mental and physical. *Philosophical Psychology*, *30*(5), 525–544.

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

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, *21*(12), 1160.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, *13*(5), 194–202.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In *Scientific explanation, space, and time*. Minneapolis: University of Minnesota Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology*, 577.

Hoffman, D. D., Prakash, C., & Prentner, R. (2023). Fusions of consciousness. *Entropy*, *25*(1), 129.

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

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, *83*(5), 291–295.

Jackson, F. (1998). Epiphenomenal qualia. In *Consciousness and emotion in cognitive science* (pp. 197–206). Routledge.

Josephs, E. L., Hebart, M. N., & Konkle, T. (2023). Dimensions underlying human understanding of the reachable world. *Cognition*, *234*, Article 105368.

Jost, J. (2015). *Mathematical concepts*. Springer.

Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., & Oizumi, M. (2023). *Is my "red" your "red"?: Unsupervised alignment of qualia structures via optimal transport*. PsyArXiv preprint.

Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., & Oizumi, M. (2023). Comparing color similarity structures between humans and LLMs via unsupervised alignment. ArXiv preprint arXiv:2308.04381.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In *Scientific explanation. Minnesota studies in the philosophy of science*. Minneapolis: University of Minnesota Press.

Kleiner, J. (2020a). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition*, *85*, Article 102981.
Kleiner, J. (2020b). Mathematical models of consciousness. *Entropy*, *22*(6), 609.
Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1), Article niab001.
Kleiner, J., & Ludwig, T. (2023). What is a mathematical structure of conscious experience? *Synthese*. In press.
Kleiner, J., & Tull, S. (2021). The mathematical structure of integrated information theory. *Frontiers in Applied Mathematics and Statistics*, *6*, 74.
Klincewicz, M. (2011). Quality space model of temporal perception. In *Multidisciplinary aspects of time and time perception* (pp. 230–245). Springer.
Kob, L. (2023). Exploring the role of structuralist methodology in the neuroscience of consciousness: A defense and analysis. *Neuroscience of Consciousness*, *2023*(1), Article niad011.
Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews. Neuroscience*, *17*(5), 307–321.
Koch, R., & Murugan, J. (2012). Emergent spacetime. In *Foundations of space and time: Reflections on quantum gravity* (pp. 164–184).
Kostic, D. (2012). The vagueness constraint and the quality space for pain. *Philosophical Psychology*, *25*(6), 929–939.
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 4.
Lau, H., Michel, M., LeDoux, J. E., & Fleming, S. M. (2022). The mnemonic basis of subjective experience. *Nature Reviews Psychology*, *1*(8), 479–488.
Lee, A. Y. (2021). Modeling mental qualities. *Philosophical Review*, *130*(2), 263–298.
Lee, A. Y. (2022). Objective phenomenology. *Erkenntnis*, 1–20.
Luce, R. D., & Suppes, P. (2004). Stevens' handbook of experimental psychology. In H. Pashler, & J. Wixted (Eds.), *Stevens' handbook of experimental psychology, methodology in experimental psychology*. John Wiley & Sons.
Lyre, H. (2022). Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1), Article niac012.
Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of Consciousness*, *2021*(2), Article niab028.
Mason, J. W. (2013). Consciousness and the structuring property of typical data. *Complexity*, *18*(3), 28–37.
Mason, J. W. (2021). Model unity and the unity of consciousness: Developments in expected float entropy minimisation. *Entropy*, *23*(11), 1444.
Metzinger, T. (1995). The problem of consciousness. In T. Metzinger (Ed.), *Conscious experience* (pp. 3–37). Imprint Academic.
Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), Article e1628.
Michel, M. (In press). The perceptual reality monitoring theory. In Herzog, M., Schurger, A., and Doerig, A. (Eds.), *Scientific Theories of Consciousness: The Grand Tour*. Cambridge University Press.
O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*(1), 127–148.
Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, *10*(5), Article e1003588.
Pashler, H., & Wixted, J. (2004). *Stevens' handbook of experimental psychology, methodology in experimental psychology, vol. 4*. John Wiley & Sons.
Prentner, R. (2019). Consciousness and topologically structured phenomenal spaces. *Consciousness and Cognition*, *70*, 25–38.
Renero, A. (2014). Consciousness and mental qualities for auditory sensations. *Journal of Consciousness Studies*, *21*(9–10), 179–204.
Resende, P. (2022). Qualia as physical measurements: A mathematical model of qualia and pure concepts. ArXiv preprint arXiv:2203.10602.
Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, *20*, 368–393.
Rosenthal, D. (2015). Quality spaces and sensory modalities. In P. Coates, & S. Coleman (Eds.), *Phenomenal qualities: Sense, perception, and consciousness* (pp. 33–65). Oxford, UK: Oxford University Press.
Rosenthal, D. M. (1991). The independence of consciousness and sensory quality. *Philosophical Issues*, *1*, 15–36.
Rosenthal, D. M. (2016). Quality spaces, relocation, and grain. In O'Shea (Ed.), *Sellars and his legacy* (pp. 149–185). Oxford: Oxford University Press.
Rovelli, C. (2004). *Quantum gravity*. Cambridge University Press.
Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, *428*, 106–131.
Safron, A. (2022). Integrated world modeling theory expanded: Implications for the future of consciousness. *Frontiers in Computational Neuroscience*, *16*, Article 642397.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Schanda, J. (2007). CIE colorimetry. *Colorimetry: Understanding the CIE System*, *3*, 25–78.
Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.
Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, *79*(7), 357–381.
Signorelli, C. M., Wang, Q., & Coecke, B. (2021). Reasoning about conscious experience with axiomatic and graphical mathematics. *Consciousness and Cognition*, *95*, Article 103168.
Signorelli, C. M., Wang, Q., & Khan, I. (2021). A compositional model of consciousness based on consciousness-only. *Entropy*, *23*(3), 308.
Silva, L. (2023). Towards an affective quality space. *Journal of Consciousness Studies*, *30*(7–8), 164–195.
Stanley, R. P. (1999). Qualia space. *Journal of Consciousness Studies*, *6*(1), 49–60.
Tallon-Baudry, C. (2022). The topological space of subjective experience. *Trends in Cognitive Sciences*.
Tononi, G. (2008). Consciousness as Integrated Information: A provisional manifesto. *Biological Bulletin*, *215*(3), 216–242.
Tononi, G. (2015). Integrated Information Theory. *Scholarpedia*, *10*(1), 4164.
Tsuchiya, N., Phillips, S., & Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition*, *101*, Article 103319.
Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), Article niab034.
Tsuchiya, N., Saigo, H., & Phillips, S. (2023). An adjunction hypothesis between qualia and reports. *Frontiers in Psychology*, *13*, Article 1053977.
Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and Integrated Information Theory. *Neuroscience Research*, *107*, 1–7.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1–24.
Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2021). *The consciousness theories studies (contrast) database: Analyzing and comparing empirical studies of consciousness theories*. biorxiv.
Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, *6*(3), 271–291.
Young, B. D., Keller, A., & Rosenthal, D. (2014). Quality-space theory in olfaction. *Frontiers in Psychology*, *5*(1).
Zaidi, Q., Victor, J., McDermott, J., Geffen, M., Bensmaia, S., & Cleland, T. A. (2013). Perceptual spaces: Mathematical structures to neural mechanisms. *The Journal of Neuroscience*, *33*(45), 17597–17602.
Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., & Tsuchiya, N. (2023). Are color experiences the same across the visual field? *Journal of Cognitive Neuroscience*, *35*(4), 509–542.

Synthese (2024) 203:89
https://doi.org/10.1007/s11229-024-04503-4

ORIGINAL RESEARCH

What is a mathematical structure of conscious experience?

Johannes Kleiner[1,2,3,4] · **Tim Ludwig**[5]

Received: 20 February 2023 / Accepted: 18 January 2024 / Published online: 5 March 2024

Abstract

Several promising approaches have been developed to represent conscious experience in terms of mathematical spaces and structures. What is missing, however, is an explicit definition of what a 'mathematical structure of conscious experience' is. Here, we propose such a definition. This definition provides a link between the abstract formal entities of mathematics and the concreta of conscious experience; it complements recent approaches that study quality spaces, qualia spaces, or phenomenal spaces; and it provides a general method to identify and investigate structures of conscious experience. We hope that ultimately this work provides a basis for developing a common formal language to study consciousness.

Keywords Quality spaces · Qualia spaces · Phenomenal spaces · Perceptual spaces · Q-spaces · Structuralism

Attempts to represent conscious experiences mathematically go back at least to 1860 (Fechner, 1860), and a large number of approaches have been developed since. They span psychophysics, philosophy, phenomenology, neuroscience, theories of consciousness, and mathematical consciousness science (Clark, 1993, 2000; Coninx, 2022; Fortier-Davy & Millière, 2020; Gert, 2017; Grindrod, 2018; Haun & Tononi, 2019; Hoffman & Prakash, 2014; Kleiner, 2020; Klincewicz, 2011; Kostic, 2012; Kuehni & Schwarz, 2008; Lee, 2021, 2022; Mason, 2013; Oizumi et al., 2014; Prent-

Johannes Kleiner
johannes.kleiner@lmu.de

1 Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

2 Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany

3 Institute for Psychology, University of Bamberg, Markusplatz 3, 96047 Bamberg, Germany

4 Association for Mathematical Consciousness Science, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

5 Institute for Theoretical Physics, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

ner, 2019; Renero, 2014; Resende, 2022; Rosenthal, 2010, 2015, 2016; Stanley, 1999; Tallon-Baudry, 2022; Tsuchiya & Saigo, 2021; Tsuchiya et al., 2022; Yoshimi, 2007; Young et al., 2014; Zaidi et al., 2013) and are known under various different names, including quality spaces (Clark, 1993), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021), Q-spaces (Chalmers & McQueen, in press), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink et al., 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020). The mathematical structures and spaces introduced by these approaches have enabled significant advancements in their respective fields. Nevertheless, this research remains largely fragmented. The various approaches employ different formalizations and different mathematical structures, and they presume a different, and sometimes partial, understanding of the concept of a mathematical structure or space when applied to conscious experience. What is missing, from our perspective, is a definition of the term 'mathematical structure of conscious experience' that clarifies how this term can and should be used.

In this article, we propose a definition of mathematical structures of conscious experience. Our main desideratum is that for a mathematical structure to be *of* conscious experience, there must be something *in* conscious experience that corresponds to that structure: a specific structural aspect of conscious experience.

Our key idea is to use variations to identify and investigate these structural aspects of conscious experience. That is because variations can serve as a binding link between conscious experiences and mathematical structures: on the one hand, variations relate to conscious experiences, because variations change aspects of conscious experiences (like qualia, qualities, or phenomenal properties); on the other hand, variations relate to mathematical structures, because they may or may not preserve them.

In defining a mathematical structure of conscious experience, our proposal does not answer the question of what this mathematical structure actually is, or which type it has. Instead, our proposal identifies the analysandum for future work on spaces and structures of conscious experience, based on which phenomenal spaces, quality spaces, qualia spaces, Φ-structures, as well as several other related concepts, can be constructed and investigated.

This paper is structured as follows. In Sect. 1, we discuss how recent approaches relate mathematical structures to conscious experience and identify three key issues in these approaches. In Sect. 2, we present our proposal together with the necessary background information. In Sects. 3, and 4, we consider two important examples: relative similarity and topological spaces. In Sect. 5, we show how our proposal resolves the three problems identified in Sect. 1. Finally, our conclusion follows in Sect. 6.

1 The status quo

So where do things stand? Most of the early work that has attributed mathematical structure to conscious experience was grounded in intuition. Whether or not a specific mathematical structure is a mathematical structure of conscious experience—a structure which "pertains to", or "belongs to" consciousness, that is—was not assessed

systematically; instead, it was assessed based on an intuitive insight of appropriateness. More recent approaches have realized the need for a more systematic method, for example Gert (2017); Lee (2021, 2023); Prentner (2019); Resende (2022); Rosenthal (2015, 2016). In this section, we analyze what we take to be the condition that underlies these approaches: a condition that justifies prescribing a mathematical structure to conscious experience. As we will see, this condition is quite natural. But, as we will demonstrate, it cannot be understood as a sufficient condition.

In a nutshell, a mathematical structure consists of two building blocks; for a detailed introduction, see Sect. 2.2. The first building block consists of one or more sets called the *domains* of the structure. The second building block are *relations or functions* which are defined on the domains. For reasons explained below, we will denote them as *structures* in the narrow sense of the term. A metric space, for example, is a mathematical structure that is defined on two domains: a set of points and the real numbers. Furthermore, it comprises a function—the so-called metric function—which maps two points to a real number. A topological space, to give another example, is a mathematical structure that is defined on a single domain: a set of points. Furthermore, it comprises a collection of unary relations, which are subsets of the domain.[1]

Usually, a mathematical structure also comes with *axioms*. The axioms establish conditions that the functions or relations have to satisfy. In the case of a metric structure, the axioms require the metric function to satisfy three conditions, called positive definiteness, symmetry, and triangle inequality. In the case of a topological structure, the axioms ensure that the collection includes the empty set and the whole domain, that it is closed under finite intersections, and that it is closed under arbitrary unions.

When put in these terms, recent proposals that go beyond intuitive assessments, make use, either directly or indirectly, of the following condition to justify that a specific mathematical structure is a mathematical structure of consciousness. Here, we use the term *aspect* as a placeholder for qualia, qualities, (instantiated) phenomenal properties, or similar concepts.[2]

(MDC) A mathematical structure is a mathematical structure of conscious experience if and only if the following two conditions are satisfied:

(D1) The domains of the structure are sets whose elements correspond to aspects of conscious experiences.

(D2) The axioms of the structure are satisfied.

In the case of the metric structure introduced in Clark (1993), for example, (D1) is satisfied because the set of points corresponds to qualities of conscious experience. The real numbers might have a phenomenal interpretation as describing degrees of similarity, as for example in Lee (2021). Condition (D2) requires positive definiteness, symmetry, and the triangle inequality to hold. This includes, for example, the condition

[1] A unary relation on a domain, in the mathematical sense, is a subset of the domain; see Sect. 4.

[2] We use the term 'aspect' as a placeholder for these terms because the above condition is not unanimously framed in either of these terms, and because our proposal in Sect. 2 is applicable with respect to any of these choices. In short, our goal is not to pick any one of these concepts but to offer a definition that works with respect to all of these concepts. Which concept is best suited for a particular task or domain is a philosophical question that can be answered independently of our proposal.

that "points should have distance zero just in case the qualities represented by those points are phenomenally identical" (Lee, 2021, p. 14). In the case of the topological structure introduced in Stanley (1999), to give another example, (D1) is satisfied because the domain of the structure refers to qualia. Condition (D2) would require, then, that the chosen collection of subsets satisfies the axioms of a topological space.

Prima facie, (MDC) could be taken to define what a mathematical structure of conscious experience is. However, if understood as sufficient condition, the following three problems arise.

Problem 1: Incompatible structures

A first reason why (MDC) cannot be a sufficient condition to assess whether a mathematical structure is a mathematical structure of consciousness is that it allows for incompatible structures.

Consider, as an example, the case of topology. A basic question in topology is whether a target domain is discrete or not. A target domain is discrete if and only if its topology contains all subsets of the domain (Joshi, 1983). Otherwise, the target domain is not discrete. These two cases are exclusive, meaning that discrete and non-discrete topological structures are incompatible.

According to (MDC), conscious experience has a discrete structure. That is because any set whatsoever can be equipped with the discrete topology. Therefore, picking a set X of aspects (qualia, qualities, phenomenal properties, etc.) and choosing its discrete topology provides a mathematical structure that satisfies both conditions (D1) and (D2). But, according to (MDC), consciousness also has a non-discrete structure. That is because any set can also be equipped with a non-discrete topology. We can, for example, take an arbitrary decomposition of the set X into two subsets A and $A^{\perp}$, where $A^{\perp}$ is the complement of A, and consider the topology $\{\emptyset, A, A^{\perp}, X\}$. This choice satisfies all axioms of a topology, and therefore satisfies (D2). Furthermore, it is built on the same set X as the discrete topology above, which implies that it also satisfies (D1). Therefore, the discrete and the non-discrete topological structures are both structures of conscious experience, according to (MDC).

This example shows that, if understood as a sufficient condition, (MDC) implies that two incompatible structures are both structures of conscious experience, and that they do so with respect to the exact same domain of aspects. The condition fails to determine which of the two incompatible structures is the right one.

Problem 2: Arbitrary re-definitions

A second reason why (MDC) cannot be a sufficient condition is that it allows for arbitrary re-definitions: if one structure is given that satisfies (MDC), then any arbitrary definition of a new structure in terms of the given structure also satisfies (MDC), so long as the domains of the structure remain unchanged. If the former pertains to consciousness, so does the latter.

A simple example of this is given by rescaling a metric function. Let us suppose that (M, d) is a metric structure which pertains to consciousness according to (MDC),

where M is a set of aspects and d is the metric function, which provides a real number $d(a, b)$ for every two aspects a and b. Since (M, d) satisfies (MDC), so does every structure $(M, C \cdot d)$, where $C \cdot d$ is the multiplication of the function d by a positive real number C. Here, the number C can be chosen arbitrarily. Therefore, if one metric structure pertains to consciousness according (MDC), so does an uncountably infinite number of metric structures.

What is more, when re-defining structures, one is free to change the axioms as one pleases. For example, we could pick any function f that maps M to the positive real numbers and define a new distance function by $(f(a) + f(b)) \cdot d(a, b)$. This might not be a metric structure anymore, because the triangle inequality axiom might not hold. But it still satisfies positive definiteness and symmetry, and therefore satisfies (MDC), with a new set of axioms. One could even break asymmetry to get a distance function like the one applied by IIT (Kleiner & Tull, 2021). More severe cases appear with more complicated structures.

This is a problem, not only because of the unlimited number of structures that appear, but also because there is an arbitrariness in the definition of a new structure, specifically concerning the axioms. It seems strange that the axioms can be redefined at will, so as to always satisfy Condition (D2). Something is missing that restricts this arbitrariness in (MDC).

Problem 3: Indifference to consciousness

The third reason, which speaks against the sufficiency of (MDC), is that the proposed condition seems somewhat indifferent to details of conscious experience.

To illustrate this indifference, let us consider again the discrete and non-discrete topological structures from above. As we have shown, these structures pertain to conscious experience according to (MDC). Yet, nothing more than a few lines needed to be said to establish this fact. In particular, we did not need to use any noteworthy input related to consciousness other than picking some set of aspects; and it didn't matter which aspects we picked.

It is a red flag if so short an analysis, which does not depend on consciousness in a meaningful way, establishes facts about the mathematical structure of conscious experience. The example exposes an indifference of (MDC) to details of conscious experience: the definition only relates to the different aspects, but not to the sort of mathematical object that connects these different aspects. Speaking somewhat vaguely, (MDC) does not refer to the "way" in which the different aspects of consciousness are related. This is why, in the case of topology, it allows one to draw conclusion without any noteworthy input from actual experience. This constitutes another reason that condition (MDC) is missing some important component, if used as sufficient condition.

Cause of these problems

These three problems arise because (MDC) is not only a necessary, but also a sufficient condition: it contains an 'if' condition in addition to the 'only if' condition. In the first

problem, we showed that two incompatible mathematical structures—a discrete and a non-discrete topology—each satisfy (D1) and (D2). Because (MDC) is a sufficient condition, it follows that both structures are structures of conscious experience, according to (MDC). In light of the incompatibility of discrete and non-discrete topologies, this constitutes an issue of the definition. In the second problem, we showed that for any given structure or space that satisfies (D1) and (D2), any arbitrary redefinition yields a structure or space which also satisfies (D1) and (D2), for a suitably adapted set of axioms. Because (MDC) is a sufficient condition, this implies that the arbitrarily redefined structure is also a mathematical structure of conscious experience, which for reasons explained above, constitutes an issue as well. The third problem, finally, built on the first one and makes use of the sufficient condition in exactly the same way. Because there is no condition in (MDC) that relates to structure in the narrow sense of the term—no condition that relates to relations or functions, that is—, and because of the sufficient condition in (MDC), structures of conscious experience can be established without reference to structure in the narrow sense of the term.

The way forward

To resolve the three problems, our task is to propose a definition for a mathematical structure of conscious experience that makes sense as a necessary and sufficient condition. This will be the content of Sect. 2.

Two desiderata guide our search. First, as is the case with (MDC), an improved definition should be *about* conscious experience in the sense that it targets qualities, qualia, instantiated phenomenal properties, or similar aspects of conscious experience, as in (D1) above. Second, there should be something *in* conscious experience—a quality, or quale or phenomenal property—that relates to structure in the narrow sense of the term. This "something" should make sure that the definition is not indifferent to conscious experience in the sense of Problem 3, and that the definition refers to functions or relations in a meaningful way, so as to stop arbitrary re-definitions (Problem 2). The proposal which we present in the next section is the result of our search.

Despite the above-mentioned problems, we think that (MDC) is an important condition. It might not be suitable as a sufficient condition, but it is valuable as a necessary condition. If one understands mathematics pragmatically as constituting a *language*—a body of symbols and terms with rules that connect these—, then mathematics can be used to describe phenomena, much like the English language can. Looking back at Condition (MDC) after our analysis, and presuming this pragmatic conception of mathematics, we think that (MDC) is best understood as an expression of what it takes for a mathematical structure to *describe* conscious experience. That is, (MDC) might be a valuable descriptive tool that utilizes mathematical structure to represent information on how aspects are related to each other (as explicated by (D1) and (D2)).

Because of this, we will refer to a mathematical structure that satisfies (MDC) as a mathematical structure that 'describes conscious experience' in what follows. The new condition which we develop below contains (MDC) as necessary part; this is aligned with the intuition that any mathematical structure of conscious experience also describes conscious experience.

2 Mathematical structures of conscious experience

In this section, we provide a definition of what mathematical structures of conscious experience are. Based on this definition, phenomenal spaces, quality spaces, qualia spaces, and related structures can be constructed and investigated. The definition embodies a way to think and work with mathematical structures when applied to conscious experience.

Our key desideratum in improving (MDC), explained above, is that for a mathematical structure to be a mathematical structure of conscious experience, rather than just a descriptive tool for conscious experience, there must be a structural aspect in conscious experience that relates to that structure. A major goal of this section is to explain this in detail. Denoting a mathematical structure by S, we call this structural aspect an S-aspect.

To make sense of what an S-aspect is, we need to understand how aspects (like qualia, qualities, or phenomenal properties) relate to mathematical structures. While aspects may have an arity, meaning they may be instantiated relative to other aspects, they are not experienced as having a mathematical structure per se (unless, of course, they are aspects of experiences of mathematical structures themselves, such as of geometric shapes). Therefore, relating aspects to mathematical structures requires a tool that applies to both: concrete aspects of conscious experience and abstract formal entities. Variations provide such a tool.

In general, a variation is a change of something into something else; in our case, it is a change of one experience into another experience. Such variations may be induced by external stimuli or interventions, occur naturally, or be subjected to imagination ('imaginary variations' (Husserl, 1936/1970)). Variations are directly related to aspects of conscious experiences because a variation can change an aspect. This is the case iff an aspect is part of the experience before the variation but isn't part of the experience after the variation. And variations are also intimately related to mathematical structures, because they may or may not preserve them, as explained in detail below. An S-aspect, then, is an aspect that is changed by a variation if and only if the variation does not preserve the structure S. To explain this in detail is the purpose of the remainder of this section.

2.1 Terminology and notation

Here, we introduce the key terms we use to define mathematical structures of conscious experience. These terms are *conscious experiences*, *aspects* of conscious experiences, and *variations* of conscious experiences. The introduction proceeds axiomatically, so that our construction does not rely on a specific choice of these concepts. Rather, any choice of these concepts that is compatible with the requirements below can be the basis of an application of our definition.

Our construction is based on a *set E of conscious experiences* of an experiencing subject. We denote individual conscious experiences in that set by symbols like e and e'; formally $e, e' \in E$. From a theoretical or philosophical perspective, one may think of the set E as comprising all conscious experiences which one experiencing

subject can have, i.e. all nomologically possible experiences of that subject. From an experimental or phenomenological perspective, one may think of this set as comprising all conscious experiences that can be induced in the lab or in introspection. Different such choices may lead to different mathematical structures being accessible.

We use the term *aspect* as a placeholder for concepts such as *qualia* (Tye, 2021), *qualities* (Clark, 1993), *mental qualities* (Rosenthal, 2010), or (instantiated) *phenomenal properties*.[3] For every experience $e \in E$, we denote the set of aspects instantiated in this experience by $A(e)$. The set of all aspects of the experiences in E, denoted by $\mathcal{A}$, is the union of all $A(e)$; formally $\mathcal{A} = \bigcup_{e \in E} A(e)$. Individual aspects, that are members of $\mathcal{A}$, will be denoted by small letters such as a, b, c. When explaining examples, we will often use the abbreviation 'a is the experience of ...' as a shorthand for saying 'a is a ... aspect of an experience'. For example, 'a is the experience of red color' means 'a is a red color aspect of an experience'.

Some aspects may require other aspects for their instantiation. For example, it is usually the case that an experience of relative similarity is an experience of relative similarity of something, for example two color aspects relative to a third color aspect. If an aspect a requires other aspects for its instantiation, we will say that the aspect a *is instantiated relative to* aspects $b_1, ..., b_m$, or simply that a *is relative to* $b_1, ..., b_m$. Aspects which are instantiated relative to other aspects are the building blocks for the structure of conscious experience.

A *variation* of a conscious experience e changes e into another experience e'. Because experiences have structure, there may be various different ways to go from e to e'.[4] Therefore, in addition to specifying e and e', a variation is a partial mapping

$$v : A(e) \to A(e') .$$

This mapping describes how aspects are replaced or reshuffled by the variation. A mapping which is not surjective, meaning that it does not map to all aspects in $A(e')$, makes room for appearance of new aspects. A mapping which is partial, meaning that it does not specify a target for every aspect in $A(e)$, makes room for aspects to disappear.

[3] Many other concepts work as well. For example, if one works with an atomistic conception of states of consciousness, where the total phenomenal state of a subject—what it is like to be that subject at a particular time—is built up from individual atomic states of consciousness, one can take e to denote the total phenomenal state and aspects to be the *states of consciousness* in that total state. Another example would be to take aspects to denote phenomenal distinctions as used in Integrated Information Theory (Tononi, 2015). What matters for our definition to be applicable is only that according to one's chosen concept of conscious experience, every conscious experience exhibits a set of aspects.

[4] To illustrate this point, consider, for example, the following two mappings v and v' which map the numbers 1, 2, and 3 to the numbers 2, 4, and 6. The mapping v is the multiplication of every number by 2, meaning that we have $v(1) = 2$, $v(2) = 4$, $v(3) = 6$. The mapping v', on the other hand, is defined by $v(1) = 6$, $v(2) = 2$, $v(3) = 4$. If we only cared about the *sets* of elements that these mappings connect, the mappings would be equivalent: there is no difference between the set $\{2, 4, 6\}$, which is the image of v, and $\{6, 2, 4\}$, which is the image of v'. If, however, we care about the *structure* of the elements of the sets—in this case, the *ordering* of numbers—, then there is a difference. While $2 \leq 4 \leq 6$, it is not the case that $6 \leq 2 \leq 4$. Because we care about the order of the elements, we need to say which element goes where.

2.2 What is a mathematical structure?

To find a rigorous definition of the mathematical structure of conscious experience, we need to work with a rigorous definition of mathematical structure. Mathematical logic provides this definition, which we now review.

A *mathematical structure* $\mathbb{S}$ consists of two things: domains, on the one hand, and functions or relations, on the other hand. We now introduce these concepts based on two simple examples.

The *domains* of a structure $\mathbb{S}$ are the sets on which the structure is built. We denote them by $\mathcal{A}_i$, where i is some index in a parameter range I. In the case of a metric structure, for example, the domains would be $\mathcal{A}_1 = M$ and $\mathcal{A}_2 = \mathbb{R}$, where M is a set of points and $\mathbb{R}$ denotes the real numbers, understood as a set. In the case of a strict partial order, there is just one domain $\mathcal{A}$, which contains the elements that are to be ordered.

The second ingredient are functions and/or relations. *Functions* f map some of the domains to other domains. In the case of a metric structure, the function would be a metric function $d : M \times M \to \mathbb{R}$, which maps from $\mathcal{A}_1 \times \mathcal{A}_1$ to $\mathcal{A}_2$. A *relation* R, in the mathematical sense, is a subset of the m-fold product $\mathcal{A}_i \times \cdots \times \mathcal{A}_i$. Here, $\mathcal{A}_i$ is the domain on which the relation is defined, and m is the arity of the relation, which expresses how many relata the relation relates. The product is usually just written as $\mathcal{A}_i^m$. In the case of a strict partial order, the relation is binary, which means that R is a subset of $\mathcal{A}^2$. For binary relations, one usually uses notation like $a < b$ instead of writing $(a, b) \in R$.

In almost all cases, mathematical structures also come with *axioms*, which establish conditions that the functions or relations have to satisfy. They are useful because they constrain and classify the structure at hand. For $\mathbb{S}$ to be a metric structure, for example, the function d has to satisfy the axioms of positive definiteness, symmetry, and triangle inequality (Rudin, 1976). For $\mathbb{S}$ to be a strict partial order, the relation R has to be irrefelxive, asymmetric, and transitive (Joshi, 1989).

To have a nice and compact notation, we will use one symbol S_j to denote both functions and relations. That is because, in any concrete proposal, it is always clear whether S_j is a function or a relation.[5] The index j takes values in some parameter range J that specifies how many functions or relations there are. Using this notation, we can represent the definition of mathematical structure provided by mathematical logic as follows:

A mathematical structure $\mathbb{S}$ is a tuple

$$\mathbb{S} = \left((\mathcal{A}_i)_{i \in I}, (S_j)_{j \in J}\right)$$

of domains $\mathcal{A}_i$ and functions or relations S_j.

[5] In mathematical logic, mathematical structures are denoted as triples of domains, relations, and functions. However, in our case, using just one symbol for functions and relations improves readability substantially.

For given domains $\mathcal{A}_i$, the mathematical structure $\mathbb{S}$ is fully determined by the S_j. Thus, we can also refer to S_j as 'structures', if the domains are clear from context. For simplicity, we can drop the index j and simply write S whenever we consider just one such structure.

As a final step in this section, we introduce the *relata* of a structure S. This will be helpful to present definitions concisely below. The term relata designates those elements that are related by a structure. In the case where S is a relation R on a domain $\mathcal{A}$ and has arity m, these are the elements of the m-tuples $(b_1, ..., b_m) \in R$. In the case where S is a function $f : \mathcal{A}_1 \times \cdots \times \mathcal{A}_{m-1} \to \mathcal{A}_m$, the relata are the elements of the m-tuples $(b_1, ..., b_{m-1}, b_m)$ where $b_m = f(b_1, ..., b_{m-1})$, and where the other b_i range over their whole domains. For notational simplicity, we write $b_1, ..., b_m$ instead of $(b_1, ..., b_m)$ when designating relata below.

2.3 What is a mathematical structure of conscious experience?

Finally, to the heart of the matter! We recall that we have so far identified two desiderata for a mathematical structure $\mathbb{S}$ to be a mathematical structure of conscious experience. First, it should be *about* conscious experiences in the sense that its domains should correspond to aspects of conscious experiences. Second, there should be aspects *in* conscious experience that relate to the structure $\mathbb{S}$. The following definition satisfies these two desiderata. Its explanation is the task of the remainder of this section.

(MSC) A mathematical structure $\mathbb{S}$ is a mathematical structure of conscious experience if and only if the following two conditions hold:

(S1) The domains $\mathcal{A}_i$ of $\mathbb{S}$ are subsets of $\mathcal{A}$.

(S2) For every S_j, there is an S_j-aspect in $\mathcal{A}$.

Here, $\mathcal{A}$ denotes the set of all aspects of the experiences in E; formally $\mathcal{A} = \bigcup_{e \in E} A(e)$, the $\mathcal{A}_i$ denote the domains of the structure $\mathbb{S}$, and the S_j-aspects are defined below.

Condition (S1) guarantees that the first desideratum is satisfied. Condition (S2) guarantees that the second desideratum is satisfied. Furthermore, whenever a certain *type* of structure (metric, topological, partial order, manifold, etc.) is claimed to be a structure of conscious experience, the axioms that constrain and classify that type have to hold. Therefore, any mathematical structure of conscious experience (MSC) is also a mathematical structure that describes conscious experience according to (MDC). The condition that has been applied in previous proposals remains a necessary condition in (MSC).

The remaining task of this section is to explain what an S_j-aspect is. For notational simplicity, we use the symbol S to denote S_j. As we have emphasized before, variations are key to understand the structure of conscious experience, because they link aspects and structure. Therefore, to be able to precisely define what an S-aspect is, we need to understand how variations relate to aspects, on the one hand, and structures, on the other hand. Our strategy is to first discuss how variations relate to aspects. This amounts to specifying what precisely it means for a variation to change an aspect.

Second, we focus on how variations relate to mathematical structure. This amounts to explaining what it means for a variation to preserve a structure. Finally, combing these two steps allows us to understand S-aspects and provide a useful definition.

What does it mean for a variation $v : A(e) \rightarrow A(e')$ to change aspects? The underlying idea is simply that an aspect is present in the source of the variation, $A(e)$, but not present any more in the target of the variation, $A(e')$. We need to take into account, though, that aspects are often instantiated relative to other aspects (see Sect. 2.1). This can be done as follows.

A variation $v{:}A(e) \rightarrow A(e')$ **changes an aspect** $a \in A(e)$ *relative to* $b_1, ..., b_m \in A(e)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(e)$, but a is not instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$.

In the case where $a \in A(e)$ is not instantiated relative to other aspects, the definition indeed reduces to the simple condition that $a \in A(e)$ but $a \notin A(e')$. The negation of the definition is also as intuitively expected: the aspect is present both in the source and in the target.[6]

For applications it is important to understand that this definition can fail to apply in two ways. First, it can fail because there is no a in $A(e')$ which is instantiated relative to $v(b_1), ..., v(b_m)$. This, in turn, can be the case either because there is no a in $A(e')$ at all, or because there is an a in $A(e')$ but it is instantiated relative to other aspects. Second, it can fail because one or more of the $v(b_1), ..., v(b_m)$ do not exist. The second case is possible because v is a *partial* mapping, which means aspects can disappear.

What does it mean for a variation to preserve a mathematical structure? The underlying idea is that a variation preserves the structure if and only if the structure is satisfied before the variation and remains to be satisfied after the variation. By its very nature, this is a mathematical condition, namely the condition of being a homomorphism (Mileti, 2022). The definition of a homomorphism, though, always applies to all elements of a domain at once. For our case, it is best to refine this definition to a single set of relata.[7]

A variation $v : A(e) \rightarrow A(e')$ **preserves a structure** S *with respect to relata* $b_1, ..., b_m \in A(e)$ if and only if we have

(P1) $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ if S is a relation R, or

(P2) $v\big(f(b_1, ..., b_{m-1})\big) = f\big(v(b_1), ..., v(b_{m-1})\big)$ if S is a function f.

[6] Because the definiendum already includes the first part of the condition, the negation is as follows: A variation $v : A(e) \rightarrow A(e')$ *does not change an aspect* $a \in A(e)$ *relative to* $b_1, ..., b_m \in A(e)$ if and only if a is instantiated relative to $b_1, ..., b_m$ in $A(e)$ and a is also instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$. We felt that is the best way of writing things to optimize clarity.

[7] For notational simplicity, we write $R\big(b_1, ..., b_m\big) = R\big(v(b_1), ..., v(b_m)\big)$ instead of $R\big(b_1, ..., b_m\big) \Leftrightarrow R\big(v(b_1), ..., v(b_m)\big)$.

As in the previous case, the negation of this definition is exactly what is intuitively expected: a variation does not preserve the structure if and only if the structure is satisfied before the variation, but not satisfied after the variation.[8]

For applications it is again important to see that the definition can fail to apply for two reasons. First, it could be the case that one or more of the $v(b_i)$ do not exist in $A(e')$, if the corresponding aspect disappears. Second, the identities may fail to hold.

We now have the keys to understand S-aspects. The underlying idea is that an S-aspect is an aspect that, under any variation, behaves exactly as the structure S does: whenever S is preserved, the S-aspect does not change, and whenever the S-aspect changes, the structure S is not preserved. This is expressed by the following definition.

An aspect $a \in \mathcal{A}$ is an S-aspect if and only if the following condition holds:
A variation *does not preserve* S with respect to relata $b_1, ..., b_m$ *if and only if* the variation *changes* a relative to $b_1, ..., b_m$.

Here, the condition needs to hold true for all variations and all relata. This means that it needs to hold true for all variations of all experiences e in the set E that instantiate relata of the structure S.

This concludes our proposal for the definition of the mathematical structure of conscious experience. It is a structure whose domains correspond to sets of aspects, and which contains an S-aspect for every relation or function of the structure. In the next two sections, we apply this definition to two examples. On the one hand, these examples illustrate the definition. On the other hand, they provide new insights to structures that have been featured prominently in previous approaches.

3 Relative similarity

Our first example concerns relative similarity, which plays an important role, for example, in the construction of quality spaces by Clark (1993, 2000).

A first step in applying our definition is to choose a set E. Here we take E to comprise experiences of three color chips, as indicated in Fig. 1A, where one of the chip (the reference) has a fixed color coating and the others vary in a range of color coatings Λ. A color coating is a physical stimuli.

The second step is to specify the set of aspects $A(e)$ for every experience $e \in E$. Here, we take $A(e)$ to comprise: (a) the color qualities in e, that is, the experienced colors of the individual chips; (b) positional qualities of the color experiences, that is, which chip has which color; and (c) the experience of *relative similarity*. Relative similarity is an experience of one pair of aspects to be more, less, or equally similar to

[8] *A variation* $v : A(e) \rightarrow A(e')$ *does not preserve a structure* S *with respect to relata* $b_1, ..., b_m \in A(e)$ if and only if we have $R(b_1, ..., b_m) \neq R(v(b_1), ..., v(b_m))$ if S is a relation R, or $v(f(b_1, ..., b_{m-1}) \neq f(v(b_1), ..., v(b_{m-1}))$ if S is a function f. This negation agrees with the intuition because the definiendum already states part of the condition that follows, namely that $b_1, ..., b_m$ are relata of the structure S in $A(e)$, which implies that $(b_1, ..., b_m) \in R$ if S is a relation and that $f(b_1, ..., b_{m-1})$ exists in $A(e)$ if S is a function, meaning that the structure is satisfied before the variation.

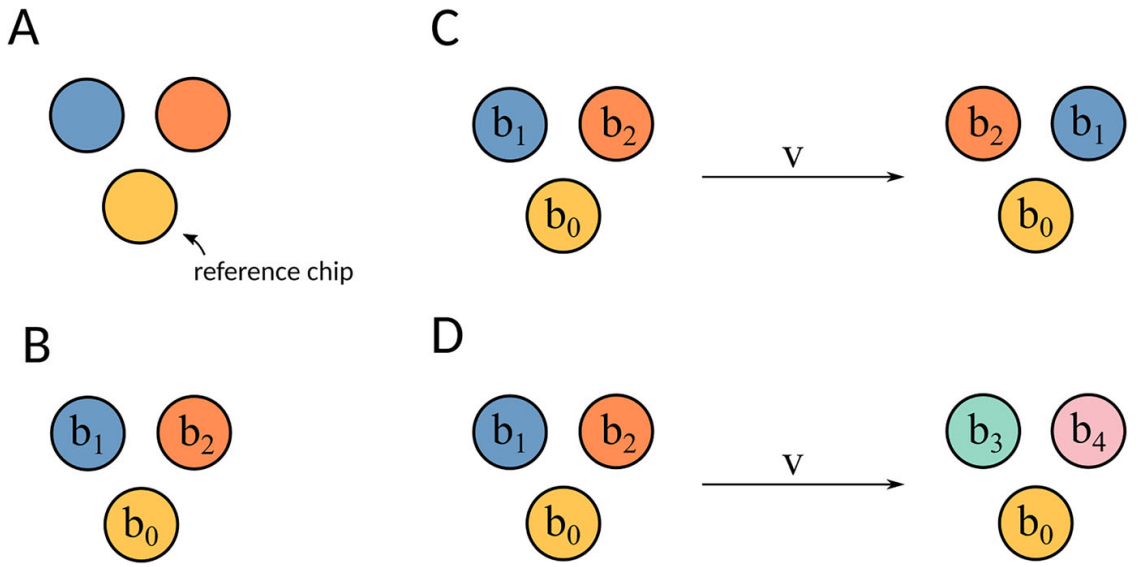


Fig. 1 To help explain the example of relative similarity, this figure illustrates experiences with color qualities and variations thereof. Subfigure **A** illustrates an experience of three color chips as well as the concept of *relative similarity*: many readers will experience the color of the top-left color chip to be *less similar* to the reference chip than the color of the top-right color chip. Subfigure **B** illustrates our notation for the color aspects corresponding to the color chips. Subfigures **C** and **D** illustrate variations v of experiences: a swap of two color aspects in **C**; and a replacement of two color aspects in **D**

each other than another pair of aspects; here, the two pairs have to have one aspect—the reference—in common. In Fig. 1A, for example, the color of the top left chip will, for many readers, be less similar to the reference chip than the color of the top right chip. An experience e in the set E may exhibit many other aspects as well. However, $A(e)$ only comprises those which are relevant for the construction at hand.

To pick out relative similarity more precisely, we let b_0, b_1 and b_2 denote the color aspects of the three chips in an experience e, where b_0 is the color aspect of the reference; see Fig. 1B. For some experience e, it might be the case that the colors b_1 and b_0 are experienced as less similar to each other than the colors b_2 and b_0. In this case, the experience e has a relative similarity aspect in the above sense; we denote this "less-similar" relative similarity aspect by a. So, a is an aspect of e, and it is instantiated relative to b_1 and b_2. (The aspect a is also relative to b_0. But since b_0 does not vary in E we can leave this implicit.)

Variations change one experience e into another experience e'. An example for a variation would be a swap of the coatings of the two non-reference chips, as in Fig. 1C. Another example for a variation would be to change the coatings of both non-reference chips to some other coating in Λ, as in Fig. 1D. Formally, variations are represented by mappings $v : A(e) \to A(e')$. In the first example, Fig. 1C, the mapping is of the form $v(b_1) = b_2$ and $v(b_2) = b_1$, and $v(c) = c$ for all other aspects c, except for the relative similarity aspect a, which is discussed in detail below. In the second example, Fig. 1D, the mapping is as in the first example but with $v(b_1) = b_3$ and $v(b_2) = b_4$.

The key question of this example is: Is there a mathematical structure of conscious experience which corresponds to relative similarity? To answer this question, we propose a mathematical structure and check whether this structure satisfies (MSC).

The words "less similar than" in the description of relative similarity already indicate that some order, in the mathematical sense of the word, might be involved. For reasons that will become clear below, we propose a strict partial order as mathematical structure. Our task in the remainder of this section is to show that this proposal indeed satisfies (MSC) with respect to experienced relative similarity. A strict partial order $(\mathcal{C}, <)$, consists of a set $\mathcal{C}$, which is the domain of the structure, and a binary relation '$<$' on $\mathcal{C}$. For all $x, y, z \in \mathcal{C}$, this binary relation has to satisfy the following axioms:

- *Irreflexivity*, meaning that there is no $x \in \mathcal{C}$ with $x < x$.
- *Asymmetry*, meaning that if $x < y$, then it is not the case that $y < x$.
- *Transitivity*, meaning that if $x < y$ and $y < z$, then also $x < z$.

In order to turn a strict partial order into a proposal for a mathematical structure of conscious experience, we need to specify how the set $\mathcal{C}$ and the relation $<$ relate to aspects of conscious experience. For the set $\mathcal{C}$ we choose the color qualities of the experiences in E, meaning that $\mathcal{C}$ now comprises the color qualities evoked by the coatings Λ of the chips we consider. For example, it contains what we have labelled b_0, b_1, b_2, b_3 and b_4 in Fig. 1. For the relation, we define $b_i < b_j$ if and only if b_i is experienced as less similar to b_0 than b_j is to b_0. (Since relative similarity, as defined above, depends on the choice of reference b_0, it would be more precise to write $<_{b_0}$ instead of $<$. However, to simplify the notation, we keep the reference implicit.)

For this proposal to make sense, we first need to check whether the axioms are satisfied. If they were not satisfied, the proposal could still be a structure of conscious experience; but it wouldn't be a strict partial order. That's why the axioms are not explicitly mentioned in (MSC). Irreflexivity is satisfied because no color quality is experienced as less similar to the reference than itself. Asymmetry is satisfied because if b_i is less similar to the reference than b_j, then b_j is not less similar to the reference than b_i.

The use of terms like 'less similar to' in natural language suggests that transitivity is also satisfied; it suggests that, if b_i is less similar to the reference than b_j and b_j is less similar to the reference than b_k, then b_i should be less similar to the reference than b_k. But it might very well be the case that natural language is not precise enough to describe its target domain. The use of natural language may be justified in simple cases, or even in a majority of cases, but whether or not transitivity holds for all $b_i, b_j, b_k \in \mathcal{C}$ is, ultimately, an empirical question. For the purpose of this example, we're going to assume that transitivity holds as well.

Having checked that the axioms hold—that is, that the proposal is indeed a strict partial order—we can proceed to check whether the structure is a mathematical structure of conscious experience according to (MSC). Concerning Condition (S1), there is one domain $\mathcal{C}$ and it consists of color qualities, so this condition is satisfied. Therefore, only Condition (S2) remains to be checked.

We now show that the relative similarity aspect a, as defined above, is in fact an S-aspect, where S is the '$<$' relation on $\mathcal{C}$. That is, it is a $<$-aspect. To see that this is true, we have to show that a variation does not preserve $<$ with respect to relata b_1 and b_2 if and only if the variation changes a relative to b_1 and b_2.

Consider any variation $v : A(e) \rightarrow A(e')$ that does not preserve $<$ with respect to relata $b_1, b_2 \in A(e)$. Two aspects b_1 and b_2 are relata of $<$ if either $b_1 < b_2$ or

$b_2 < b_1$. We focus on the first case as the other one follows from the first by renaming b_2 and b_1 in what follows. By definition of the $<$ relation, $b_1 < b_2$ means that b_1 is experienced as less similar to the reference than b_2. Therefore, there is also a relative similarity aspect $a \in A(e)$ as defined above. As explained in Sect. 2.3, there can be two ways in which the variation v might not preserve $<$. Either $v(b_1)$ or $v(b_2)$ are not defined, or, if they are defined, it is not the case that $v(b_1) < v(b_2)$. In the former case, there cannot be an a in $A(e')$ relative to $v(b_1)$ or $v(b_2)$, simply because the latter do not both exist. In the latter case, it follows from the definition of $<$ that $v(b_1)$ is not experienced as less similar to the reference than $v(b_2)$. So, there is no $a \in A(e')$ relative to $v(b_1)$ and $v(b_2)$. Hence, we may conclude that v changes a relative to b_1 and b_2.

For the other case, let $v : A(e) \to A(e')$ be a variation which preserves $<$ with respect to relata b_1 and b_2. As before, this implies that a is in $A(e)$ relative to b_1 and b_2. Because v preserves $<$, $v(b_1)$ and $v(b_2)$ both exist and we also have $v(b_1) < v(b_2)$. Applying the definition of $<$ then implies that a is also in $A(e')$ relative to $v(b_1)$ and $v(b_2)$. Hence v does not change a relative to b_1 and b_2.

Because in both of these cases, v was arbitrary, it follows that a is indeed a $<$-aspect. Therefore, Conditions (S1) and (S2) of (MSC) are both satisfied, and the strict partial order $(\mathcal{C}, <)$ is indeed a mathematical structure of conscious experience; it is the mathematical structure of relative similarity of color experiences with respect to b_0.

4 Phenomenal unity and topological structure

Our second example concerns topological structure. Interestingly, this is intimately tied to phenomenal unity, the thesis that phenomenal states of a subject at a given time are unified (Bayne & Chalmers, 2003). Phenomenal unity gives rise to a mathematical structure of conscious experience.[9]

Recall that we have introduced the set $A(e)$ to denote aspects of the conscious experience e, where we have used the term 'aspect' as a placeholder for concepts like qualia, qualities, or (instantiated) phenomenal properties. Most examples of these concepts are "independent" from the experience in which they occur; they could be experienced together with a largely different set of aspects in a different experience. Yet, experiences seem unified; their aspects are experienced as tied together in some essential way. This raises the question of what underlies this experience of the *unity of a conscious experience*? As we will see, somewhat surprisingly, the answer is: a topological structure of conscious experience.

Much has been written about the question of phenomenal unity in the literature, for example Bayne (2012); Bayne and Chalmers (2003); Cleeremans and Frith (2003); Mason (2021); Prentner (2019); Roelofs (2016); Wiese (2018), and in order to make

[9] A connection between topology and phenomenal unity has already been conjectured in Prentner (2019), where an attempt was made to construct a topological space based on a binary relation that describes the "overlap" of mental objects. The construction only leads to the weaker notion of a pre-topology, but should be regarded as an important first step in this direction. For a summary of the formal construction, see Kleiner (2020, Example 3.22).

use of some of the results, we assume that the term 'aspect' denotes an instantiated phenomenal property or quale. The set of aspects $A(e)$, then, comprises the phenomenal properties or qualia which are instantiated in the experience e, also called the *phenomenal states* of the experience e.[10] Our question, then, is what it means that "any set of phenomenal states of a subject at a time is phenomenally unified" (Bayne & Chalmers, 2003, p. 12).

There are various answers one might give to this question. A promising answer is the so-called *subsumptive unity thesis*, developed in Bayne and Chalmers (2003):

> For any set of phenomenal states of a subject at a time, the subject has a phenomenal state that subsumes each of the states in that set. (Bayne & Chalmers, 2003, p. 20)

According to this thesis, what underlies the experience of the unity of a conscious experience is that for any set X of phenomenal states in the conscious experience, there is a further phenomenal state that subsumes each of the states in X. This phenomenal state characterizes what it is like to be in all of the states of X at once (Bayne & Chalmers, 2003, p. 20).

Put in terms of aspects, the subsumptive unity thesis says that for any set $X \subset A(e)$ of aspects of an experience, there is an additional aspect in $A(e)$ that subsumes the aspects in X. This aspect is the experience of what it is like to experience the aspects in X as part of one experience e together; this aspect is the experience that the aspects in X are *unified*, as we will say. Let us call this aspect the *phenomenal unity aspect* of X and denote it by a_X. It is instantiated relative to the elements of X.

Phenomenal unity gives rise to a mathematical structure of conscious experience. To see how, let us use the symbol $\mathcal{T}$ to denote a collection of subsets of $A(e)$, to be specified in more detail below. Every subset of $A(e)$ is a unary relation on $A(e)$,[11] and hence also on the set $\mathcal{A}$ that comprises all aspects of the experiences in E. Therefore, $(\mathcal{A}, \mathcal{T})$ is a mathematical structure; it has the domain $\mathcal{A}$ and its structures are the unary relations in $\mathcal{T}$. As we show next, because of the subsumptive unity thesis, the mathematical structure $(\mathcal{A}, \mathcal{T})$ is a mathematical structure of conscious experience according to (MSC).

Because $\mathcal{A}$ is the set of all aspects of E, Condition (S1) of (MSC) is satisfied. Therefore, only Condition (S2) remains to be checked. This condition is satisfied because for every set $X \in \mathcal{T}$, the phenomenal unity aspect a_X is an S-aspect for $S = X$; an X-aspect for short. To show that this is the case, we need to check that a variation does not preserve X with respect to relata $b_1, ..., b_m$ if and only if it changes a_X relative to $b_1, ..., b_m$. Let $v : A(e) \rightarrow A(e')$ be a variation that does not preserve X with respect to relata $b_1, ..., b_m$. The relata of the subset X are the elements of that subset. Therefore, we have $b_1, ..., b_m \in A(e)$, so that the subsumptive unity thesis implies that there is a phenomenal unity aspect a_X relative to the $b_1, ..., b_m$ in $A(e)$.

[10] A *phenomenal state* is an instantiation of a phenomenal property, or quale, by a subject at a given time. This instantiation constitutes part of the experience of the subject at the time. An experience e, in our terminology, is an experience of a subject at a given time. Hence, a phenomenal state is an instantiation of a phenomenal property, or quale, in an experience e.

[11] An m-ary relation on a set X is a subset R of X^m. Hence, a unary relation, where $m = 1$, is a subset of X.

The condition that v does not preserve X furthermore implies that either not all of the $v(b_i)$ exist or that at least one of them is not in the set X. Therefore, there is no phenomenal unity aspect a_X relative to $v(b_1), ..., v(b_m)$ in $A(e')$. Hence, the variation v changes a_X relative to $b_1, ..., b_m \in X$. Vice versa, let $v : A(e) \rightarrow A(e')$ be a variation which preserves X with respect to relata $b_1, ..., b_m$. This implies that a_X is instantiated relative to $b_1, ..., b_m$ in $A(e)$. The condition that v preserves X furthermore implies that $v(b_1), ..., v(b_m)$ exist, and that they are elements of X. Therefore, a_X is also instantiated relative to $v(b_1), ..., v(b_m)$ in $A(e')$. This shows that the variation does not change a_X relative to $b_1, ..., b_m$. Thus, a_X is indeed an X-aspect. And because that is true for any $X \in \mathcal{T}$, $(\mathcal{A}, \mathcal{T})$ indeed satisfies Condition (S2) and hence (MSC).

The previous paragraph proves that, if the subsumptive unity thesis holds true for all sets X in $\mathcal{T}$, then $(\mathcal{A}, \mathcal{T})$ is indeed a mathematical structure of conscious experience. As we will explain next, this structure is intimately tied to a topological structure.

A topological structure $(M, \mathcal{T})$ consists of a set M and a collection $\mathcal{T}$ of subsets of M. The collection has to satisfy three axioms, and there are a few different ways of formulating these axioms. Here, we choose the formulation that corresponds to what is usually called 'closed sets'. The axioms are:

- The empty set $\emptyset$ and the whole set M are both in $\mathcal{T}$.
- The intersection of any collection of sets of $\mathcal{T}$ is also in $\mathcal{T}$.
- The union of any finite number of sets of $\mathcal{T}$ is also in $\mathcal{T}$.

Are these axioms satisfied by the structure $(\mathcal{A}, \mathcal{T})$ induced by phenomenal unity?

To answer this question, it is important to note that the subsumptive unity thesis does not provide a phenomenal unity aspect a_X for every subset of $\mathcal{A}$. It can only provide such an aspect for a set of aspects that are actually experienced together. That is, it can only provide such an aspect for a subset X of $A(e)$. Therefore, $\mathcal{T}$ is not the discrete topology introduced in Sect. 1. Second, it also cannot be the case that it provides a phenomenal unity aspect for every subset of $A(e)$. That's because then there would be an infinite regress: for every subset X of $A(e)$ there would be a new aspect a_X in $A(e)$, giving a new subset $X \cup \{a_X\}$ that would give a new phenomenal unity aspect $a_{X \cup \{a_X\}}$, and so forth. This problem is well-known in the literature (Bayne, 2005; Wiese, 2018). Rather, we take it, the quantifier 'any set' in the subsumptive unity thesis must be understood as 'any set of aspects that are experienced as being unified'. While it is arguably the case that the whole set of aspects $A(e)$ of an experience is always experienced as unified—by which we mean: the whole set of aspects is experienced—, introspection suggests that we consciously experience only a select group of aspects as unified at a time.[12]

So, which sets of aspects do we experience as unified? While we cannot give a general answer to this question here, there is a special case where a sufficiently detailed specification can be given: the case of regions in visual experience. Here, 'regions' are

[12] This solves the infinite regress problem because, arguably, we do not always experience the phenomenal unity aspects as unified with the sets they correspond to. So, there is not always a phenomenal unity aspect $a_{X \cup \{a_X\}}$ for the set that consists of a_X and X.

sets of positions of the space that visually perceived objects occupy.[13] The positions in a region are experienced as unified. Therefore, the regions of visual experience are members of the collection $\mathcal{T}$ which is induced by phenomenal unity. Furthermore, they appear to satisfy the axioms of a topology as stated above: the whole set of positions in a visual experience is a region; it seems to be the case that intersections of regions in visual experience are also regions in visual experience; and it seems to be the case that the union of any two regions in visual experience is also a region in visual experience. For the empty set, no S-aspect of consciousness is required (there are no relata of the corresponding unary relation), so we may take the empty set to be a member of $\mathcal{T}$. Thus, all axioms of a topology are satisfied.

Therefore, if we take M to denote the position aspects of visual experiences, and choose $\mathcal{T}$ to comprise the regions of visual experience, then $(M, \mathcal{T})$ is indeed a topological structure. And, as shown above, it is a structure of conscious experience as defined in (MSC). We thus find that, because of the subsumptive unity thesis, this topological structure is indeed a mathematical structure of conscious experience; much like conjectured in Tallon-Baudry (2022), it is a topology of the visual content of subjective experience.

5 The three problems revisited

In this section, we discuss how the new approach (MSC), which we have developed in Sect. 2.2, resolves the three problems discovered in Sect. 1.

Problem 1: Incompatible structures

The first problem was that the condition (MDC), which has been applied in previous approaches, admits incompatible structures to conscious experience. Is this also true of (MSC)?

If two structures are incompatible, then there exists at least one automorphism of one structure that is not an automorphism of the other structure.[14] As we explain below, this condition implies that two incompatible structures cannot have all S-aspects in common. Therefore, it is not possible for two incompatible structures to pertain to conscious experience in the exact same way; so, (MSC) indeed resolves the problem of incompatible structures.

Let S and S' denote two incompatible structures (in the narrow sense of the term) with the same domains. Then, there is at least one automorphism of one structure that is not an automorphism of the other structure. Let us denote such an automorphism

[13] It is also plausible to think that visual experiences do not contain positions as aspects, but only regions. However, assessing whether or not this is the case goes beyond the scope of this paper. Here, we assume that positions are aspects of visual experiences.

[14] Automorphisms are structure-preserving mappings from a structure to itself. Put in terms of the terminology we have introduced in Sect. 2.2, automorphisms are mappings v that map the domains of a structure to themselves. These mappings have to be bijective, and they have to preserve the structure, meaning that they have to satisfy (P1) for all elements of the domain in case of relations, and (P2) for elements of the domains in the case of functions.

by v and assume that it is an automorphism of S but not of S'. Because v is not an automorphism of S', it follows that there is at least one set of relata $b_1, ..., b_m$ of S' in some $A(e)$, such that the variation $v : A(e) \rightarrow A(e)$ induced by the automorphism does not preserve S' with respect to these relata. On the other hand, because v is an automorphism of S, it follows that this variation preserves S with respect to $b_1, ..., b_m$. If an aspect a is an S'-aspect, then, applying the definition of S'-aspects, we find that the variation v needs to change it. In contrast, if an aspect a is an S-aspect, then, applying the definition of S-aspects, we find that the variation v must not change it; either because the $b_1, ..., b_m$ do not constitute relata of S, or because the variation v preserves S with respect to relata $b_1, ..., b_m$. Because an aspect cannot be both changed and not changed under a single variation, there cannot be an aspect a that is both an S-aspect and an S'-aspect.

Problem 2: Arbitrary re-definitions

The definition (MSC) also resolves the problem of arbitrary re-definitions. That's because any re-definition changes the relations or functions of the respective structure, and therefore generates an own, independent condition for something to be an S-aspect of the redefined structure. Whether or not this new S-aspect is a part of conscious experience is a substantive question that depends on the actual experiences of the subject under consideration; it is not automatically the case.

Consider, as examples, the cases of rescaling a metric, which we have introduced in Sect. 1. If, per assumption, (M, d) were a structure of conscious experience, then for any relata $(b_1, b_2, d(b_1, b_2))$, the condition for d-aspects would have to be satisfied. Rescaling this to $(M, C \cdot d)$ generates a new condition because now, the relata to be considered are $(b_1, b_2, C \cdot d(b_1, b_2))$. These are different relata, and correspondingly, different experiences and different variations will enter the definition of a $C \cdot d$-aspect. The same is true for an $(f(a) + f(b)) \cdot d(a, b)$-aspect. Whether or not these structures satisfy (MSC) depends on the details of the conscious experiences under consideration; but they do not automatically satisfy (MSC) just because (M, d) does.

Problem 3: Indifference to consciousness

The third problem is resolved, finally, because of the introduction of S-aspects in (MSC), which are a counterpart "in" conscious experience to the structure in the narrow sense of the term. S-aspects introduce a connection between functions or relations in a mathematical structure, on the one hand, and aspects (qualia, qualities, or phenomenal properties) of conscious experiences, on the other hand. Because S-aspects are part of the definition of (MSC), any application of (MSC) requires engaging with details of the conscious experiences of the subject under consideration; (MSC) is not indifferent to conscious experience in the sense of Problem 3 of Sect. 1.

Consider, for example, the two topological structures of Sect. 1. While (MDC) only required us to check whether the structures address aspects and satisfy the axioms, (MSC) also requires us to check whether there is an S-aspect in conscious experience that corresponds to the topological structures. As we have seen in Sect. 4, this involves

a careful investigation of conscious experience and relies on intricate notions such as phenomenal unity.

6 Conclusion

In this article, we investigated mathematical structures and mathematical spaces of conscious experience. We were not concerned with questions of type or explicit form of these structures or spaces, but with the question of what it means to speak about mathematical structures or mathematical spaces of conscious experiences in the first place. We answer this question by providing a definition of what *mathematical structures of conscious experience* are. This definition provides a foundation for the construction, investigation and identification of concepts like phenomenal spaces, quality spaces, qualia spaces and Q-structures.

Our definition of mathematical structures of conscious experiences is grounded in a foundational understanding of mathematical structures and spaces as laid out by mathematical logic. And it is axiomatic in the sense that it can be applied to any conceptualization of conscious experiences, and any choice of aspects thereof (e.g. qualia, qualities, phenomenal properties, phenomenal distinctions), which satisfy the formal requirement that for every conscious experience there is a well-defined set of aspects.

Our definition rests on the notion of *variations*, which are changes of one conscious experience to another. Because variations can be induced introspectively (for example, as in Husserl's imaginary variations), stimulated in a laboratory by change of stimuli, or studied theoretically based on a proposed theory of consciousness, our definition constitutes a general method to identify and study structures of conscious experience.

The grounding of mathematical structures of conscious experiences proposed here is *methodologically neutral* in the sense that it can be combined with many methods, practices, and procedures that are used to investigate conscious experience, spanning empirical, analytical, and phenomenological research. Furthermore, it is *conceptually neutral* in the sense that it can be applied to any conception of 'conscious experience' and 'aspects' thereof, as long as every conscious experience comes with a well-defined set of aspects. This includes common conceptions using qualities, qualia, or phenomenal properties, but also less common ideas based on atomistic conceptions of states of consciousness or phenomenal distinctions.

Our definition complements recent approaches that study quality spaces, qualia spaces, or phenomenal spaces, because it retains the abstract condition that these proposals apply—Condition (MDC) in Sect. 1—as a necessary part. This abstract condition is extended by our proposal, so as to avoid three problems that interfere with recent approaches, see Sect. 1.

In light of the increasing interest in using mathematical structures to model and represent conscious experiences in the scientific study of consciousness and philosophy of mind, the investigation of how to define and understand mathematical structures of conscious experience is important, in our view. This work contributes to this investigation. It highlights issues with previous ways of understanding structural claims and offers an improved conception that rests on meaningful desiderata. Hence, we

hope, it contributes to building a foundation for structural research for both theory and experimental practice.

As a first application, and to illustrate our definition, we considered *relative similarity* and *topological spaces*. We found that relative similarity, which plays an important role in several constructions of quality spaces, is indeed a mathematical structure of conscious experience, see Sect. 3. Topological spaces also qualify as mathematical structures of conscious experience, but for a surprising reason: they are intimately related to phenomenal unity, see Sect. 4.

We view the results presented here as one further step in a long journey to investigate conscious experience mathematically. This step raises new questions and creates new opportunities, both of which can only be explored in an interdisciplinary manner. A new question, for example, is whether our result on mathematical structures might open new perspectives on measurements of consciousness (Irvine, 2013), as arguably promised by the Representational Theory of Measurement (Krantz et al., 1971) whenever an axiomatic structure on a target domain is available. We hope that, ultimately, our result provides a basis for developing a common formal language to study consciousness across domains.

Acknowledgements We would like to thank the participants of the *2022 Modelling Consciousness Workshop* and of the *Models of Consciousness 3* conference, both organized under the umbrella of the Association for Mathematical Consciousness Science, as well as members of the *Munich Center for Mathematical Philosophy* for fruitful discussions and helpful comments, and in particular Jonathan Mason, Robin Lorenz, Ian Durham, Christian List, Andrew Lee, Etienne Jacques and Pedro Resende for valuable feedback on the manuscript. This research was supported by grant number FQXi-RFP-CPW-2018 from the Foundational Questions Institute and Fetzer Franklin Fund, a donor advised fund of the Silicon Valley Community Foundation. We would like to thank the Dutch Research Council (NWO) for (partly) financing TL's work on project number 182.069 of the research programme Fluid Spintronics, and the Mathematical Institute of the University of Oxford for hosting JK while working on this project.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that there are no financial or non-financial competing interests that are directly or indirectly related to the work submitted for publication.

References

Bayne, T. (2005). Divided brains and unified phenomenology: A review essay on Michael Tye's consciousness and persons. *Philosophical Psychology, 18*(4), 495–512.

Bayne, T. (2012). *The unity of consciousness*. Oxford University Press.

Bayne, T. J., & Chalmers, D. J. (2003). What is the unity of consciousness? In A. Cleeremans (Ed.), *The unity of consciousness.* Oxford University Press.
Chalmers, D. J., & McQueen, K. J. (in press). Consciousness and the collapse of the wave function. In S. Gao (Ed.), *Consciousness and quantum mechanics*. Oxford University Press.
Clark, A. (1993). *Sensory qualities*. Clarendon Library of Logic and Philosophy.
Clark, A. (2000). *A theory of sentience*. Clarendon Press.
Cleeremans, A., & Frith, C. (2003). *The unity of consciousness*. Oxford University Press.
Coninx, S. (2022). A multidimensional phenomenal space for pain: Structure, primitiveness, and utility. *Phenomenology and the Cognitive Sciences, 21*(1), 223–243.
Fechner, G. (1860). *Elements of psychophysics* (Vol. I). Holt, Rinehart and Winston.
Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences, 2*. https://doi.org/10.33735/phimisci.2021.79
Fortier-Davy, M., & Millière, R. (2020). The multi-dimensional approach to drug-induced states: A commentary on Bayne and Carter's "dimensions of consciousness and the psychedelic state". *Neuroscience of Consciousness, 2020*(1), niaa004.
Gert, J. (2017). Quality spaces: Mental and physical. *Philosophical Psychology, 30*(5), 525–544.
Grindrod, P. (2018). On human consciousness: A mathematical perspective. *Network Neuroscience, 2*(1), 23–40.
Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy, 21*(12), 1160.
Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology, 5*, 577.
Husserl, E. (1936/1970). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy*. Northwestern University Press.
Irvine, E. (2013). Measures of consciousness. *Philosophy Compass, 8*(3), 285–297.
Joshi, K. (1983). *Introduction to general topology*. Wiley Eastern.
Joshi, K. D. (1989). *Foundations of discrete mathematics*. New Age International.
Kleiner, J. (2020a). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition, 85*, 102981.
Kleiner, J. (2020b). Mathematical models of consciousness. *Entropy, 22*(6), 609.
Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness, 2021*(1), niab001.
Kleiner, J., & Tull, S. (2021). The mathematical structure of integrated information theory. *Frontiers in Applied Mathematics and Statistics, 6*, 74.
Klincewicz, M. (2011). Quality space model of temporal perception. In *Multidisciplinary aspects of time and time perception* (pp. 230–245). Springer.
Kostic, D. (2012). The vagueness constraint and the quality space for pain. *Philosophical Psychology, 25*(6), 929–939.
Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. I–III). Academic Press.
Kuehni, R. G., & Schwarz, A. (2008). *Color ordered: A survey of color systems from antiquity to the present*. Oxford University Press.
Lee, A. Y. (2021). Modeling mental qualities. *Philosophical Review, 130*(2), 263–298.
Lee, A. Y. (2022). Objective phenomenology. *Erkenntnis,* 1–20. https://doi.org/10.1007/s10670-022-00576-0
Lee, A. Y. (2023). Degrees of consciousness. *Noûs*. *Noûs, 57*, 553–575. https://doi.org/10.1111/nous.12421
Lyre, H. (2022). Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness, 2022*(1), niac012.
Mason, J. W. (2013). Consciousness and the structuring property of typical data. *Complexity, 18*(3), 28–37.
Mason, J. W. (2021). Model unity and the unity of consciousness: Developments in expected float entropy minimisation. *Entropy, 23*(11), 1444.
Mileti, J. (2022). *Modern mathematical logic*. Cambridge University Press.
Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology, 10*(5), e1003588.
Prentner, R. (2019). Consciousness and topologically structured phenomenal spaces. *Consciousness and Cognition, 70*, 25–38.
Renero, A. (2014). Consciousness and mental qualities for auditory sensations. *Journal of Consciousness Studies, 21*(9–10), 179–204.

Resende, P. (2022). Qualia as physical measurements: A mathematical model of qualia and pure concepts. arXiv preprint arXiv:2203.10602
Roelofs, L. (2016). The unity of consciousness, within subjects and between subjects. *Philosophical Studies, 173*(12), 3199–3221.
Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues, 20*, 368–393.
Rosenthal, D. (2015). Quality spaces and sensory modalities. In P. Coates & S. Coleman (Eds.), *Phenomenal qualities: Sense, perception, and consciousness* (pp. 33–65). Oxford University Press.
Rosenthal, D. M. (2016). Quality spaces, relocation, and grain. In O'Shea (Ed.), *Sellars and his legacy* (pp. 149–185). Oxford University Press.
Rudin, W. (1976). *Principles of mathematical analysis* (Vol. 3). McGraw-Hill.
Stanley, R. P. (1999). Qualia space. *Journal of Consciousness Studies, 6*(1), 49–60.
Tallon-Baudry, C. (2022). The topological space of subjective experience. *Trends in Cognitive Sciences, 26*(12), 1068–1069.
Tononi, G. (2015). Integrated information theory. *Scholarpedia, 10*(1), 4164.
Tsuchiya, N., Phillips, S., & Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition, 101*, 103319.
Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness, 2021*(2), niab034.
Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/qualia/
Wiese, W. (2018). What is it like to experience a third man? The phenomenological Bradley and how to solve it. In W. Wiese (Ed.), *Experienced wholeness: Integrating insights from Gestalt theory, cognitive neuroscience, and predictive processing*. The MIT Press. https://doi.org/10.7551/mitpress/9780262036993.003.0004
Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences, 6*(3), 271–291.
Young, B. D., Keller, A., & Rosenthal, D. (2014). Quality-space theory in olfaction. *Frontiers in Psychology, 5*, 1.
Zaidi, Q., Victor, J., McDermott, J., Geffen, M., Bensmaia, S., & Cleland, T. A. (2013). Perceptual spaces: Mathematical structures to neural mechanisms. *Journal of Neuroscience, 33*(45), 17597–17602.

Title: The Case for Neurons: A No-Go Theorem for Consciousness on a Chip

Johannes Kleiner, University of Bamberg, Institute for Psychology
Tim Ludwig, Utrecht University

Acknowledgements

I am deeply grateful to all who have made this thesis possible. First and foremost, my supervisor Stephan Hartmann. Thank you, Stephan, for the guidance, inspiration, and support you have so kindly provided. Second, my collaborators, whom I would like to thank for countless hours of discussion on the topic we all love so much. The research presented here would not have been possible without you. Third, my academic friends and colleagues, both locally and internationally, for the inspiration and questions you provide, for pushing ahead with some of the best research I know, for creating a field that is so pleasant to work in, and for advice and guidance. And fourth, I would like to thank my non-academic friends and especially my family for supporting me in my academic journey despite the uncertainties that it brings with it. Thank you all!

I would like to thank the *Graduate School of Systemic Neurosciences* and the *Munich Center for Mathematical Philosophy*, both at *Ludwig Maximilian University of Munich*, for hosting and supporting this PhD thesis. It was a great pleasure to be part of the fruitful intellectual atmosphere that both institutions provide. Furthermore, I would like to thank the *Mathematical Institute* of *Oxford University* as well as the *NYU Center for Mind, Brain, and Consciousness* of *New York University* for hosting me while working on some of the projects presented here.

Author Contributions

Author contributions are specified using the Contributor Roles Taxonomy **CRediT** (Brand, Allen, Altman, Hlava, & Scott, 2015), defined as follows (CRediT taxonomy, 2021):

- **Conceptualization:** Ideas; formulation or evolution of overarching research goals and aims.
- **Data curation:** Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.
- **Formal analysis:** Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data.
- **Funding acquisition:** Acquisition of the financial support for the project leading to this publication.
- **Investigation:** Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
- **Methodology:** Development or design of methodology; creation of models.
- **Project administration:** Management and coordination responsibility for the research activity planning and execution.
- **Resources:** Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.
- **Software:** Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
- **Supervision:** Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.
- **Validation:** Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
- **Visualization:** Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.
- **Writing – original draft:** Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
- **Writing – review & editing:** Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

Publication 1: Kleiner, J., & Tull, S. (2021). The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, 6, 602973.

Johannes Kleiner, Sean Tull

Author contributions:

J.K.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

S.T.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Publication 2: Tull, S., & Kleiner, J. (2021). Integrated Information in Process Theories: Towards Categorical IIT. *Journal of Cognitive Science*, 22, 2, 92–123.

Sean Tull, **Johannes Kleiner**

Author contributions:

S.T.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

J.K.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Writing – review & editing.

Publication 3: Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, 2021(1), niab001.

Johannes Kleiner, Erik Hoel

Author contributions:

J.K.: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

E.H.: Conceptualization, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing.

Publication 4: Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, 119, 103653.

Johannes Kleiner

Author contributions:

J.K.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Publication 5: Kleiner, J., & Ludwig, T. (2024). What is a mathematical structure of conscious experience?. *Synthese*, 203(3), 89.

Johannes Kleiner, Tim Ludwig

Author contributions:

J.K.: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

T.L.: Conceptualization, Investigation, Methodology, Validation, Writing – review & editing.

Publication 6: Kleiner, J., & Ludwig, T. (2024). The Case for Neurons: A No-Go Theorem for Consciousness on a Chip. Forthcoming in *Neuroscience of Consciousness*.

Johannes Kleiner, Tim Ludwig

Author contributions:

J.K.: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing.

T.L.: Conceptualization, Investigation, Methodology, Validation, Writing – review & editing.