Three Comments in Case of a Structural Turn in Consciousness Science

Johannes Kleiner[1,2,3,4]

[1]Munich Center for Mathematical Philosophy, LMU Munich
[2]Munich Graduate School of Systemic Neurosciences, LMU Munich
[3]Institute for Psychology, University of Bamberg
[4]Association for Mathematical Consciousness Science

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 structuralist approaches, overlooked assumptions in regard to isomorphisms, and the question of what structure to consider on the side of consciousness 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—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 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 [25], 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" [38].

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 pioneering work by Austen Clark [7] and David Rosenthal [58], mathematical spaces are now applied in philosophy [8, 9, 13, 15, 43, 44, 56, 57, 59, 11, 45, 37, 53, 53, 80, 6, 68, 2], neuroscience [72, 70, 82, 42, 46, 17, 52, 19, 28, 75, 83, 18], cognitive science [24, 61, 23, 51], psychology [36, 40, 54, 81] and mathematical consciousness science [16, 32, 69, 55, 47, 48, 66, 76, 74, 73, 31, 33, 34]. They are known under various different names, including quality spaces [7, 57], qualia spaces [69], experience spaces [33, 35, 56], Q-spaces [6, 45], Q-structure [45], Φ-structures [72], perceptual spaces [82], phenomenal spaces [11], spaces of subjective experience [70], and spaces of states of conscious experiences [31]. The first theory of consciousness to make use of mathematical spaces was IIT 2.0 [71]; more recent versions expand and refine the idea [52, 1].

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

[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.

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 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 38 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 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 of how a brain mechanism, or physical system more generally, relates to a mathematical space or mathematical structure. Because a mathematical space or structure consists of a set plus some relation or function on

[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: 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, 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 set, there is much more information to provide in specifying a space or structure than specifying only a set. Furthermore, this information has to satisfy constraints to provide a legitimate definition of a mathematical object, such as the axioms of a particular space. So 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.

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) [1] and Expected Float Entropy Minimisation Theory (EFE) [48]. While both theories represent significant advances, establishing a link to existing phenomenal spaces (cf. Section 5) remains a next-level challenge.[3]

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.

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,[4] and the new predictions about structural facts might offer new avenues for experimental investigation.[5]

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 [11]. 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

[3]Proponents of both theories are fully aware of this task, and IIT has made a first step in this direction in [17].

[4]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 for more on that distinction.

[5]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* [79].

discuss and criticise the key assumption that enables this methodology—the assumption of a homomorphism between phenomenal and neuronal structures—in Section 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" [50] of sorts.

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.[6]

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 out one's location in the wilderness without a map or other 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. That's possible because maps store information about geography. Another example of this sort is quantum tomography, where a

[6]I'm grateful to a conversation with Lucia Melloni about the problem of reports and similar measures during a walk at the above-mentioned *Structuralism in Consciousness Studies* workshop.

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-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 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 their 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 [27, 26], and for modal arguments like colour inversion [65, 4].

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 [44].

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 [63], 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.

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 be applied independently of metaphysical assumptions, and in research programs of any metaphysical flavour. Structural approaches do not have metaphysical premises, and they do not come with a preferred metaphysical interpretation.

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 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 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 of a Riemannian manifold. If we talk about the structure of a neural network, we mean the mathematical structure of the graph that specifies the connectivity of the network.

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 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. 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.* [7] Vice versa, if the automorphism group of a structure is trivial, then every point must have a unique location.[8]

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, we therefore find that 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:

[7] 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.

[8] 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.

(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.[9]

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. 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 making use of structures (in the general sense of mathematical structure) and structuralist assumptions, 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. 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" [5], 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. 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. 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 [10]. If epistemic phenomenal structuralism is true, then one cannot know 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,

[9]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.

if private or intrinsic properties exist (or if there are properties which are not accessible to structural cognitive processing, for that matter), 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, [32], 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." [29].

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" [11], 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 [11, 45], might even "open an attractive door for reductionism" [11].

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 as understood in the context of structuralism in consciousness science. 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. [2], 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 [66] and [67] 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 independently 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*, 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 map more generally. 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.

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. Strictly speaking, though, homomorphisms are not appropriate either,[10]

[10] For the nerds ;-) 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: 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

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.[11]

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 [25], 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" [37].

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) [52, 1]. 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" [52]. The mathematical algorithm of the theory specifies a mapping

multiple realizability, the 'isomorphism onto the image' intuition seems to be closer to the underlying intuition. The same applies if one reverses the direction of the homomorphism, cf. Footnote 11.

[11] In addition to the problem mentioned in Footnote 10, 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 [11], 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 10 always exists, so that the statement becomes empty. 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.

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.[12]

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* [64], 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 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

[12]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" [52], and because the mathematical quantities outputted by IIT's algorithm do not appear anywhere else in the physical sciences, and are conceptually and mathematically rather involved, the choice of regarding this structure as physical seems somewhat arbitrary, to say the least. It 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 an empty 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).

thought. They only map points in one domain to points in another domain in such a way that the mapping between the points *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 [21, 20] to more modern understandings of explanation in the form of causal-mechanical models [62], unificationist models [14, 30], contrastive explanation [77] or interventionalist models [78, 22].

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 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.

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, 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.

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.[13]

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.

[13]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.

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 what one takes phenomenal structure to denote. 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.[14]

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 [7], quality spaces as introduced by David Rosenthal [58, 56] and Q-spaces as introduced in IIT 2.0 [71]. 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" [52]. This gives rise to what David Chalmers has called the *Rosetta Stone Problem* [5]: the problem of how to translate the mathematical structure that IIT proposes into phenomenological terms. In other words, 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.

[14]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'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 [60] and emergent spacetime [39] to proposals that depart completely from what we intuitively think spacetime should be [12]. 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" [49]. 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.

3. **Conscious Experience.** Only part of the mathematical structure actually pertains to conscious experiences or phenomenal character.

Clark's Quality Spaces. Quality spaces as introduced by Austen Clark [7] are based on the following methodology. To construct the quality space for an individual subject,[15] 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 *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*.[16]

The discrimination task is used to define a *global indiscriminability* relation on the class of stimuli $\mathcal{S}$.[17] 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 terminology) 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. 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 and the labels represent relative similarity. Mathematically speaking, it is a POSET-labelled graph, where POSET means '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 tasks should depend on what is experienced.

[15] Clark mostly has humans in mind, but does consider the case of animals briefly in [7]. Nothing hinges on humans in the methodology he proposes.

[16] 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.

[17] 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 $\mathcal{S}$.

The next step in Clark's methodology consists of introducing a metric, and in fact a Euclidean space. To this end, it makes use of a procedure known as 'multidimensional scaling' [3]. 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 stored on a computer in familiar ways. In either case, the last step which 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. And 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.[18] 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

[18]For a more careful examination of the case of a metric, cf. [34]. For questions on how quality spaces *should* relate to consciousness or phenomenal character according to the underlying theory, cf. below.

a subject notices a difference—requires a *topology* on the stimulus space. The mathematical structure of topology is precisely what provides continuous curves in this sense. A topology defines what it means to "draw a line without lifting a pen" on an abstract space, so to speak. 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 behind the methodology which can be used to answer these and similar worries. When I asked David Rosenthal about the problem regarding variations, for example, he countered by assuming that there is just one actual physical topology in reality and this is what should be used.[19] Similarly, there is a theory that discharges the methodology from the problem that, according to the usual understanding of the discrimination task in this case, discriminations could also be made unconsciously.

It is good that these and other problems of Rosenthal's methodology have thorough answers if one presumes his theoretical point of view. But if quality spaces are to furnish a structural turn to the extent indicated in Section 2, they cannot presume a specific theory. This is why I have always been tempted, maybe mistakenly, 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. Section 5 and [34]).

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 experiences*, to use a general term. In a nutshell, all existing proposals I know of amount to:

(A) Conditions on the domains 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.

[19] It's not clear to me how this would work in practice, I should add, given that this topology is presumably defined by Quantum Electrodynamics (QED). The topology of QED states is too far removed from experimental practice to be applicable, on my view.

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 [34, 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 [34], Tim Ludwig and I have set out to find a methodology that achieves this task. Our proposal shares with David Rosenthal's methodology that it rests on variations, though in our case, any transition from one conscious experience to another conscious experience counts as variation, we do not demand continuity or restrict to changes that are induced by changes 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.

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 [41], 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 methodologies 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.

Acknowledgments. I’d like to take this opportunity to thank the organisers and participants of the recent *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. 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 for hosting me while working on this comment.

References

[1] L. Albantakis, L. Barbosa, G. Findlay, M. Grasso, A. M. Haun, W. Marshall, W. G. Mayner, A. Zaeemzadeh, M. Boly, B. E. Juel, et al. Integrated information theory (iit) 4.0: formulating the properties of phenomenal existence in physical terms. *arXiv preprint arXiv:2212.14787*, 2022.

[2] H. Atmanspacher. The Pauli–Jung conjecture and its relatives: A formally augmented outline. *Open Philosophy*, 3(1):527–549, 2020.

[3] R. Beals, D. H. Krantz, and A. Tversky. Foundations of multidimensional scaling. *Psychological review*, 75(2):127, 1968.

[4] N. Block. Inverted earth. *Philosophical perspectives*, 4:53–79, 1990.

[5] D. Chalmers. Phenomenal structuralism. Talk presented at the Structuralism in Consciousness Studies Workshop at the Charité Berlin, 2023.

[6] D. J. Chalmers and K. J. McQueen. Consciousness and the collapse of the wave function. In S. Gao, editor, *Consciousness and Quantum Mechanics*. Oxford University Press, forthcoming.

[7] A. Clark. *Sensory qualities*. Clarendon Library of Logic and Philosophy, 1993.

[8] A. Clark. *A theory of sentience*. Clarendon press, 2000.

[9] S. Coninx. A multidimensional phenomenal space for pain: structure, primitiveness, and utility. *Phenomenology and the Cognitive Sciences*, 21(1):223–243, 2022.

[10] D. C. Dennett. Quining qualia. *Consciousness in contemporary science*, pages 42–77, 1988.

[11] S. B. Fink, L. Kob, and H. Lyre. A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, 2, 2021.

[12] F. Finster and J. Kleiner. Causal fermion systems as a candidate for a unified physical theory. In *Journal of Physics: Conference Series*, volume 626, page 012020. IOP Publishing, 2015.

[13] M. Fortier-Davy and R. Millière. 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, 2020.

[14] M. Friedman. Explanation and scientific understanding. *the Journal of Philosophy*, 71(1):5–19, 1974.

[15] J. Gert. Quality spaces: Mental and physical. *Philosophical Psychology*, 30(5):525–544, 2017.

[16] P. Grindrod. On human consciousness: A mathematical perspective. *Network neuroscience*, 2(1):23–40, 2018.

[17] A. Haun and G. Tononi. Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, 21(12):1160, 2019.

[18] J.-D. Haynes. Decoding visual consciousness from human brain signals. *Trends in cognitive sciences*, 13(5):194–202, 2009.

[19] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

[20] C. G. Hempel. Deductive-nomological vs. statistical explanation. In *Scientific explanation, space, and time*. University of Minnesota Press, Minneapolis, 1962.

[21] C. G. Hempel and P. Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.

[22] C. Hitchcock and J. Woodward. Explanatory generalizations, part ii: Plumbing explanatory depth. *Noûs*, 37(2):181–199, 2003.

[23] D. D. Hoffman and C. Prakash. Objects of consciousness. *Frontiers in Psychology*, page 577, 2014.

[24] D. D. Hoffman, C. Prakash, and R. Prentner. Fusions of consciousness. *Entropy*, 25(1):129, 2023.

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

[26] F. Jackson. What mary didn't know. *The journal of philosophy*, 83(5):291–295, 1986.

[27] F. Jackson. Epiphenomenal qualia. In *Consciousness and emotion in cognitive science*, pages 197–206. Routledge, 1998.

[28] E. L. Josephs, M. N. Hebart, and T. Konkle. Dimensions underlying human understanding of the reachable world. *Cognition*, 234:105368, 2023.

[29] J. Jost. *Mathematical concepts*. Springer, 2015.

[30] P. Kitcher. Explanatory unification and the causal structure of the world. In *Scientific explanation. Minnesota studies in the philosophy of science*. University of Minnesota Press, Minneapolis, 1989.

[31] J. Kleiner. Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition*, 85:102981, 2020.

[32] J. Kleiner. Mathematical models of consciousness. *Entropy*, 22(6):609, 2020.

[33] J. Kleiner and E. Hoel. Falsification and consciousness. *Neuroscience of Consciousness*, 2021(1):niab001, 2021.

[34] J. Kleiner and T. Ludwig. What is a mathematical structure of conscious experience? *arXiv preprint arXiv:2301.11812*, 2023.

[35] J. Kleiner and S. Tull. The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, 6:74, 2021.

[36] M. Klincewicz. Quality space model of temporal perception. In *Multidisciplinary aspects of time and time perception*, pages 230–245. Springer, 2011.

[37] L. Kob. Exploring the role of structuralist methodology in the neuroscience of consciousness: a defense and analysis. *Neuroscience of Consciousness*, 2023(1):niad011, 2023.

[38] C. Koch, M. Massimini, M. Boly, and G. Tononi. Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17(5):307–321, 2016.

[39] R. Koch and J. Murugan. Emergent spacetime. *Foundations of space and time: reflections on quantum gravity*, pages 164–184, 2012.

[40] D. Kostic. The vagueness constraint and the quality space for pain. *Philosophical Psychology*, 25(6):929–939, 2012.

[41] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

[42] H. Lau, M. Michel, J. E. LeDoux, and S. M. Fleming. The mnemonic basis of subjective experience. *Nature Reviews Psychology*, 1(8):479–488, 2022.

[43] A. Y. Lee. Modeling mental qualities. *Philosophical Review*, 130(2):263–298, 2021.

[44] A. Y. Lee. Objective phenomenology. *Erkenntnis*, pages 1–20, 2022.

[45] H. Lyre. Neurophenomenal Structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, 2022(1):niac012, 2022.

[46] R. Malach. Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of consciousness*, 2021(2):niab028, 2021.

[47] J. W. Mason. Consciousness and the structuring property of typical data. *Complexity*, 18(3):28–37, 2013.

[48] J. W. Mason. Model unity and the unity of consciousness: Developments in expected float entropy minimisation. *Entropy*, 23(11):1444, 2021.

[49] T. Metzinger. The problem of consciousness. In T. Metzinger, editor, *Conscious experience*, pages 3–37. Imprint Academic, 1995.

[50] M. Michel. Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, 14(2):e1628, 2023.

[51] G. O'Brien and J. Opie. A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, 22(1):127–148, 1999.

[52] M. Oizumi, L. Albantakis, and G. Tononi. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, 2014.

[53] R. Prentner. Consciousness and topologically structured phenomenal spaces. *Consciousness and Cognition*, 70:25–38, 2019.

[54] A. Renero. Consciousness and mental qualities for auditory sensations. *Journal of Consciousness Studies*, 21(9-10):179–204, 2014.

[55] P. Resende. Qualia as physical measurements: a mathematical model of qualia and pure concepts. *arXiv preprint arXiv:2203.10602*, 2022.

[56] D. Rosenthal. How to think about mental qualities. *Philosophical Issues*, 20:368–393, 2010.

[57] D. Rosenthal. Quality spaces and sensory modalities. In P. Coates and S. Coleman, editors, *Phenomenal qualities: sense, perception, and consciousness*, pages 33–65. Oxford University Press Oxford, UK, 2015.

[58] D. M. Rosenthal. The independence of consciousness and sensory quality. *Philosophical issues*, 1:15–36, 1991.

[59] D. M. Rosenthal. Quality spaces, relocation, and grain. In O'Shea, editor, *Sellars and his Legacy*, pages 149–185. Oxford University Press Oxford, 2016.

[60] C. Rovelli. *Quantum gravity*. Cambridge university press, 2004.

[61] D. Rudrauf, D. Bennequin, I. Granic, G. Landini, K. Friston, and K. Williford. A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, 428:106–131, 2017.

[62] W. C. Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[63] J. Schanda. Cie colorimetry. *Colorimetry: Understanding the CIE system*, 3:25–78, 2007.

[64] A. Seth. *Being you: A new science of consciousness*. Penguin, 2021.

[65] S. Shoemaker. The inverted spectrum. *The Journal of Philosophy*, 79(7):357–381, 1982.

[66] C. M. Signorelli, Q. Wang, and B. Coecke. Reasoning about conscious experience with axiomatic and graphical mathematics. *Consciousness and Cognition*, 95:103168, 2021.

[67] C. M. Signorelli, Q. Wang, and I. Khan. A compositional model of consciousness based on consciousness-only. *Entropy*, 23(3):308, 2021.

[68] L. Silva. Towards an affective quality space. *Journal of Consciousness Studies*, 30(7-8):164–195, 2023.

[69] R. P. Stanley. Qualia space. *Journal of Consciousness Studies*, 6(1):49–60, 1999.

[70] C. Tallon-Baudry. The topological space of subjective experience. *Trends in Cognitive Sciences*, 2022.

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

[72] G. Tononi. Integrated Information Theory. *Scholarpedia*, 10(1):4164, 2015.

[73] N. Tsuchiya, S. Phillips, and H. Saigo. Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition*, 101:103319, 2022.

[74] N. Tsuchiya and H. Saigo. A relational approach to consciousness: categories of level and contents of consciousness. *Neuroscience of Consciousness*, 2021(2):niab034, 2021.

[75] N. Tsuchiya, H. Saigo, and S. Phillips. An adjunction hypothesis between qualia and reports. *Frontiers in Psychology*, 13:1053977, 2023.

[76] N. Tsuchiya, S. Taguchi, and H. Saigo. Using category theory to assess the relationship between consciousness and Integrated Information Theory. *Neuroscience research*, 107:1–7, 2016.

[77] B. C. Van Fraassen. *The scientific image*. Oxford University Press, 1980.

[78] J. Woodward and C. Hitchcock. Explanatory generalizations, part i: A counterfactual account. *Noûs*, 37(1):1–24, 2003.

[79] I. Yaron, L. Melloni, M. Pitts, and L. Mudrik. The consciousness theories studies (contrast) database: analyzing and comparing empirical studies of consciousness theories. biorxiv, 2021.06. 10.447863, 2021.

[80] J. Yoshimi. Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, 6(3):271–291, 2007.

[81] B. D. Young, A. Keller, and D. Rosenthal. Quality-space theory in olfaction. *Frontiers in Psychology*, 5:1, 2014.

[82] Q. Zaidi, J. Victor, J. McDermott, M. Geffen, S. Bensmaia, and T. A. Cleland. Perceptual spaces: mathematical structures to neural mechanisms. *Journal of Neuroscience*, 33(45):17597–17602, 2013.

[83] A. Zeleznikow-Johnston, Y. Aizawa, M. Yamada, and N. Tsuchiya. Are color experiences the same across the visual field? *Journal of Cognitive Neuroscience*, 35(4):509–542, 2023.