NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2022, **2022(1)**, 1–18

DOI: https://doi.org/10.1093/nc/niac012

Research Article

Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness

Holger Lyre*,†

Chair for Theoretical Philosophy, University of Magdeburg, Magdeburg, Germany;
Center for Behavioral Brain Sciences (CBBS), University of Magdeburg, Magdeburg, Germany;
Research Training Group 2386 "Extrospection", Humboldt-Universität zu Berlin, Berlin, Germany
†Holger Lyre, http://orcid.org/0000-0002-6040-0263
*Correspondence address. Lehrstuhl für Theoretische Philosophie, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, Magdeburg D-39106, Germany.
Email: lyre@ovgu.de

Abstract

The program of "neurophenomenal structuralism" is presented as an agenda for a genuine structuralist neuroscience of consciousness that seeks to understand specific phenomenal experiences as strictly relational affairs. The paper covers a broad range of topics. It starts from considerations about neural change detection and relational coding that motivate a solution of the Newman problem of the brain in terms of spatiotemporal relations. Next, phenomenal quality spaces and their Q-structures are discussed. Neurophenomenal structuralism proclaims a homomorphic mapping of the structures of self-organized neural maps in the brain onto Q-structures, and it will be demonstrated how this leads to a new and special version of structural representationalism about phenomenal content. A methodological implication of neurophenomenal structuralism is that it proposes measurement procedures that focus on the relationships between different stimuli (as, for instance, similarity ratings or representational geometry methods). Finally, it will be shown that neurophenomenal structuralism also has strong philosophical implications, as it leads to holism about phenomenal experiences and serves to reject inverted qualia scenarios.

Keywords: quality spaces; structural similarity; Newman problem; structural representation; self-organized neural maps; neurophenomenal holism; structural qualia; qualia inversion

Introduction

What is the relation of phenomenal states to neural states? And how do they relate to the external world of stimuli? These are the kinds of questions I want to address in this paper. And I try to answer them within a philosophical program called "neurophenomenal structuralism" that I intend to develop here. The paper is largely of programmatic nature and builds on ideas already spelled out in a joint paper with Sascha Fink and Lukas Kob (Fink *et al.* 2021, henceforth cited as FKL). Neurophenomenal structuralism, as I develop it here, can indeed be seen as an agenda for a genuine structuralist neuroscience of consciousness.

In FKL, we declared that neurophenomenal structuralism "rests on the idea that (i) each experience is, in scientific contexts, structurally individuated and that (ii) there is a systematic relation between phenomenal structures and structures in the neural domain" (FKL, p. 3). We then went on to claim that "a strong metaphysical reading of these twin ideas is possible: Every phenomenal character is exhaustively individuated by its relations and not, as qualia theorists hold, by any intrinsic phenomenal properties. We are sympathetic to this reading, but focus here on a methodological point" (FKL, p. 3). Complementary to the FKL paper, the focus of my present paper is on ontology. More properly, the focus is on a neuroscientifically informed ontology. I try to explore in the most general way possible the broad range of insights and philosophical implications that follow from the idea that "phenomenal character is exhaustively individuated by its relations." Hence, I go well beyond the FKL paper. But the starting point is the same. The starting point and the spirit of neurophenomenal structuralism can also be found in the writings of other authors.[1] But ultimately, the position of neurophenomenal structuralism is independent. It should in particular become clear that the position, as I develop it here, must be understood as a rigorous structuralist ontology about phenomenal consciousness that is physicalist and reductionist in spirit. It can also be seen as part of the bigger project of a general structuralist conception of the mind. In the course of the paper, three broad domains and their associated structures

[1] First and foremost, the work of Austen Clark (1993, 2000) must be mentioned. Stronger or weaker connections can be drawn (no claim to completeness) to the writings of O'Brien and Opie (1999, 2001, 2004), Chalmers (2012), Gert (2017a, 2017b), Isaac (2013, 2014), Loorits (2014), and Rosenthal (2010, 2015) in philosophy, Palmer (1978, 1999), Shepard (1968, 1970), and Edelman (1998) in psychology, and, more recently, Tsuchiya and Saigo (2021) and Malach (2021) in neuroscience.

Received 3 March 2022; **Revised** 4 July 2022; **Accepted** 27 July 2022

shall be distinguished, and it will be helpful to use abbreviations for them:

- The W-domain of worldly states and processes (external stimuli)
- The Q-domain of qualitative phenomenal states and processes (phenomenal experiences)
- The N-domain of neural states and processes (neural activities)

The three domains can be described or modeled in terms of the "state spaces" that are spanned by the domain states. For W, this is a physical state space, the particular nature of which is of no concern for this paper. For Q, these are the various spaces of phenomenal states or "quality spaces."[2] Finally, for N, the relevant state spaces are the activation spaces of neural networks as well as the so-called neural maps. The Q- and N-spaces will be discussed in part 2 of the paper. As the term neurophenomenal structuralism already indicates, the focus will be on the relevant structure of the three domains and their state spaces. More precisely, neurophenomenal structuralism deals with the question of how phenomenal structure, the structure of the Q-domain (or Q-structure, for short) relates to neural structure, the structure of the N-domain (N-structure). More rigorous definitions of the notion of structure and structural mappings shall be given in the "Structure, homomorphisms, and structural individuation" section, but as a first informal definition, we may think of a structure as a set of relations defined over a set of relata. The relata, in turn, are not specified otherwise than by the relations in which they stand. Neurophenomenal structuralism is now based on the following two assumptions:

1. Any phenomenal experience is fully individuated by its place in a Q-structure.
2. Q-structure is mirrored in N-structure.

The crucial ontological implication of the first assumption is that "phenomenal experience is a relational, not an intrinsic affair." It is the purpose of the first part of the paper to motivate this claim bottom-up. The notion of mirroring in the second assumption should be understood in terms of structural similarity. Therefore, we shall consider structure-preserving maps. More precisely, in the section "N-structure: neural activation spaces and neural maps," I will argue for structural mappings from N to Q in terms of surjective homomorphisms.

Neurophenomenal structuralism, in its attempt to clarify the structure of phenomenal experiences and the relation to their underlying neural structures, aims to contribute to the neuroscience of consciousness. But it should be clarified from the outset that it is about specific rather than generic consciousness. It does not tell us when a mental state is conscious in general, but rather "what the specific content of phenomenally conscious states is." And neurophenomenal structuralism's answer to the specificity question is mainly entailed in the first assumption, which, as I shall argue in the section "Phenomenality, content, and characters," leads to the same metaphysical implication for both phenomenal content and character: they are rooted in relational, not in intrinsic properties (in contrast to the traditional notion of qualia). Also, since neurophenomenal structuralism characterizes phenomenal experience essentially by content, the account turns out as a modest variant of structural representation, albeit of a very special sort (as will be laid out in the section "Structural representation").

The organization of the paper is as follows. It consists of three parts plus introduction and conclusion. The first part deals with key preliminary conceptual considerations for the justification of neurophenomenal structuralism. This concerns the nature of the neural/non-neural interface (see the "The neural/non-neural interface" section), the concepts of change detection and relational coding (see the "Change detection and relational coding" section), the Newman problem of the brain and a proposal for its solution (see the section "The Newman problem of the brain and a proposal for its solution"), and sensory intersections (see the "Sensory intersections" section). The second part serves to introduce the relevant concepts and definitions of neurophenomenal structuralism. After presenting the central technical notions (see the "Structure, homomorphisms, and structural individuation" section), the concepts of phenomenal content and character will be discussed (see the "Phenomenality, content, and characters" section). Next, phenomenal quality spaces will be introduced (see the "Quality spaces, Q-structure and psychophysics" section). Thereafter, the concepts of neural activation spaces and neural maps will be assessed (see the "N-structure: neural activation spaces and neural maps" section). The part concludes with an evaluation of the available empirical evidence for structuralism and the prospects for the neuroscience of consciousness (see the "Empirical evidence and consciousness studies" section). In the third part, the philosophical implications of neurophenomenal structuralism will be explored. It is clarified that it is not a version of structural realism (see the "Structural realism" section) but corresponds to reductive physicalism (see the "Reductive physicalism" section) and the concept of structural representation (see the "Structural representation" section). It is finally argued that neurophenomenal structuralism leads to neurophenomenal holism (see the "Neurophenomenal holism" section) and serves to reject inverted qualia scenarios (see the "Qualia inversion and (finally) the hard problem" section).

The paper concludes with a comprehensive summary in the form of a list of bullet points containing the main assertions and conclusions of the various sections. Readers may use this list already while reading the text for an overview and as a guideline.

From the external world into the neural system (and back)

This first part of the paper deals with key preliminary conceptual considerations for the justification of neurophenomenal structuralism. This comprises the nature of the neural/non-neural interface (see the section "The neural/non-neural interface"), the concepts of change detection and relational coding (see the section "Change detection and relational coding"), the Newman problem of the brain and a proposal for its solution in terms of spatiotemporal relations (see the section "The Newman problem of the brain and a proposal for its solution"), and sensory intersections (see the section "Sensory intersections").

The neural/non-neural interface

Let us start as general as possible. Our world consists of either neural or non-neural systems. They connect at a specific interface that I will call the neural/non-neural ("3N") interface or boundary. The 3N boundary consists of two parts.

The first part may be called the "boundary of neural transduction." Here, the nature of the incoming signals, i.e. the nature of

[2] Note that the notation has been changed compared to the FKL paper. There, the letter P has been used for the phenomenal. In the philosophy of mind literature, however, P often stands for the physical, while Q is used for the phenomenal (as indicating qualitative phenomenal experience).

the distal causes that affect the neural system at its sensory surfaces, changes. This change is crucial. It means that, whatever the nature of the distal cause perturbing the neural system was (be it, for instance, of electromagnetic, chemical, or acoustic nature) gets translated into neural activity and spike trains inside the neural system. Conversely speaking, this means that from the perspective of the neural system, the nature of the distal perturbing causes remains hidden. Or so it seems.

The reverse of neural transduction is the transformation of efferent motor neuron activity into bodily effector organ actions, mainly the action of muscles and glands. This is the second part of the 3N interface. Call it the "boundary of neuromotor transformation." Taken together, the two parts combine to the full neural/non-neural boundary. It, again, consists of the inward part of neural transduction and the outbound part of neuromotor transformation.

The 3N boundary should not be confused with the boundary between N and W. The reason is that the states and processes of the neural not only exclude the states and processes of W but also the states and processes of the body. On this count, W comprises the domain of states and processes beyond skin and scull. The "B-domain" of states and processes of the body is thus "sandwiched" between N and W. This is just a convention that I have chosen in this paper. I could have attributed B to W, which would then have led to a somewhat unusual usage of the terms "world" and "external stimuli." For the present purposes, however, neither would make much difference, since we are dealing here with external stimuli in the conventional sense of being outside skin and skull. In other words, we are dealing with exteroception. But this is not to say that the B-domain is of no importance for the neuroscience of consciousness. On the contrary, it is precisely for this reason that it makes sense to terminologically separate the B-domain from the W-domain. For reasons of space, however, this separation does not play a role in the present paper. It will be a task for the future to think about neurophenomenal structuralism in the context of, on the one hand, the various "internal" stimuli and signals that emanate from within the body falling under the (not consistently defined) terms interoception, proprioception, visceroception, etc. (cf. Ritchie and Carruthers 2015) as well as, on the other hand, the programs of embodied and situated cognition (or 4E cognition in general; cf. Robbins and Aydede 2009, Lyre 2013; Newen *et al.* 2018).[3]

Change detection and relational coding

Given the 3N boundary, the intrinsic nature of the external signals cannot be transferred from the external world into the neural system. This is first and foremost an "ex negativo" characterization of the 3N boundary. Positively speaking, then, what can get transferred?

Let us, again, be as general as possible. Whenever the neural system gets perturbed, i.e. whenever the system undergoes some change at its sensory periphery, that change can potentially be detected. For instance, a photon of a certain energy hits a photoreceptor in the retina and induces a change in the receptor membrane, which ideally and via a cascade of further changes in the downstream cells (the bipolar, horizontal or amacrine cells) leads to a spike, or, more likely, a train of spikes, in one or more retinal ganglion cells.

What is crucial in the above is the term "change." Neural systems, at the boundary of neural transduction, are sensitive to changes in the environmental stimuli. Neural sensory systems may "detect" such changes. Perception, in essence, works by "change detection." To say that perception works by change detection is to say that only causal perturbations or changes at the boundary of neural transduction can be transferred beyond that boundary.

From an ontological perspective, the notion of change detection takes the following form: Regardless of whether the causal stimuli of the world can be assigned intrinsic or relational properties, only relational properties will be transferred. It is fairly easy to see what these relational properties on the level of the neural system amount to, namely either differences in the intensity or activation rate or differences in the temporal structure of the neuronal action potentials or spikes. In other words, the internal neural system works on a purely relational basis: the relational properties of neuronal activities. These general considerations apply regardless of the particular neural coding scheme that is used by the system, be it rate coding, temporal coding, or population coding. Any of the neural coding schemata deliver what we may most generally call "difference or relational coding." Difference coding is solely based on the relational properties of the coding elements.

It is perhaps worthwhile to add that these considerations apply rather straightforwardly to the recent program of predictive coding or predictive processing (cf. Hohwy 2013; Clark 2016). Here, the idea is that rather than considering the perceiving brain as a passive, stimulus-driven feature detection machinery, predictive processing models complement these bottom-up processes with ongoing top-down predictions on the basis of generative models (on many levels of the cortical hierarchy). What gets processed and will eventually be represented in the neural code is the error signal, i.e. the difference between bottom-up input and top-down predictions. Predictive processing, therefore, rather naturally embraces difference coding.

The Newman problem of the brain and a proposal for its solution

Before we can go on to unfold neurophenomenal structuralism based on change detection and relational coding in part 2, we have to make a detour first over a stumbling obstacle that seems to lie in our way. It is, in fact, a rather generic problem for structuralist conceptions: the so-called Newman problem. I propose a way to overcome this problem based on spatiotemporal relations (and enriched by sensory intersections, see the section "Sensory intersections").

According to the received view, neural systems deploy neural representations to represent (aspects of) the world. Hence, the W-N-relationship is a relationship of representation. However, in order to represent, the N-representations must somehow refer to W, i.e. they must somehow cross the 3N boundary. Yet, because of that boundary, the two domains N and W seem to be clearly separated. It also seems to be the case that no direct reference can be maintained between N and W. And the problem already starts with

[3] From a statistical or machine learning point of view the 3N boundary can be viewed as a Markov blanket (cf. Pearl 1988; Friston *et al.* 2021). A Markov blanket defines the boundary of a system in a statistical sense, i.e. in terms of a partitioning of the system into internal and external variables. The blanket itself consists of the variables separating the two such that the external variables are conditionally independent of the internal ones and vice versa. By considering the variables as nodes or states in a (Bayesian) network, the Markov blanket of a node or state contains the node's or state's causal parents, children and children's parents. The Markov blanket thus covers or shields the internal states from the external states. And this is precisely how the 3N boundary acts (with afferent states of the boundary of neural transduction as parental for the internal states of the neural system and efferent states of the boundary of neuromotor transformation as children). Note, however, that it is not the mere fact of representing a Markov blanket that makes the 3N boundary ontologically distinct. Indeed, Bruineberg *et al.* (2021) have recently criticized the idea to reify Markov blankets. The ontological distinctiveness of 3N lies in the nature of the neural system.

sensory perception. How do perceptual states refer to their external causes? Upon a structural conception of N, as articulated here, this directly leads to the Newman problem.

The problem consists in the following: Given a set of entities that are not determined in any further way beyond their mere cardinality, this set can be endowed with any set of relations (i.e. structure) that is compatible with the cardinality of the set of entities. Or, in the words of Max Newman who raised this as an objection against Russell's (1927) early version of structural realism (cf. section "Structural realism"):

> …it is meaningless to speak of the structure of a mere collection of things […] Further, no important information about the aggregate A, except its cardinal number, is contained in the statement that there exists a system of relations, with A as field, whose structure is an assigned one. For given any "aggregate" A, a system of relations between its members can be found having any assigned structure compatible with the cardinal number of A. (Newman 1928, 140)

Here is a simple example: Consider a finite set A of n "naked" objects (i.e. nothing is determined beyond the mere cardinality n). Now let A be equipped with an order relation that we may indicate by the symbol ">". What can be known about ">"? Next to nothing, indeed. Although it is tempting to interpret ">" as "greater than" (as we are used to read the symbol this way), there is no reason to do so. It could likewise be interpreted in any other arbitrary sense such as "older than," "more jealous than" or "possessing 5417 protons more than." The reason for this freedom of choice is of course that neither the nature of the objects nor the nature of the relation has been determined. In fact, it could be any order relation compatible with the cardinality.[4] As Newman puts it: "… 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. So structuralism is near-vacuous, in effect it collapses to empiricism. All we can know is just cardinality."

The point of the Newman problem is not only that relations do not suffice to pick out the intrinsic nature of the objects in the domain but that also the nature of the relations themselves is indetermined.[5] As far as we have introduced and discussed N-structure, the perceiving and thinking subject seems to be threatened by the Newman problem. From the point of view of the neural system, it seems that nothing can be "known" about the external W-causes perturbing N. Is there any rescue to the Newman problem? How could we avoid it? How did Russell avoid it (if at all)? Again, Newman had raised his objection in a 1928 article on Russell's "Analysis of Matter" (Russell 1927). In this work, Russell had spelled out a rather strong version of structural realism:

> …whenever we infer from perceptions it is only structure that we can validly infer; […] and structure is what can be expressed by mathematical logic. The only legitimate attitude about the physical world seems to be one of complete agnosticism as regards all but its mathematical properties. (Russell 1927, 254, 270)

Newman's objection had an immediate impact on Russell. In April 1928 he sent a letter to Newman:

> Dear Newman, […] It was quite clear to me, as I read your article, that I had not really intended to say what in fact I did say, that nothing is known about the physical world except its structure. I had always assumed spatio-temporal continuity with the world of percepts, that is to say, I had assumed that there might be co-punctuality between percepts and non-percepts. … And co-punctuality I regarded as a relation which might exist among percepts and is itself perceptible.

According to Russell and framed in his own terminology: the answer to Newman is that we are "directly acquainted" with certain spatiotemporal relations. And this, I think, is indeed the route that we must take. Perception is based on change detection and neural systems work on a purely relational basis. What is transmitted over the 3N boundary are relational properties only. The question is, whether the nature of (some of the) relations can be conveyed as well. On the face of it, the answer seems to be negative, as all W-relations are transformed into N-relations. The crucial point, however, is that certain spatiotemporal W-relations do indeed carry over to N-relations. We can indeed directly refer to certain spatiotemporal W-relations—or, in Russell's words, are "directly acquainted" with them.

Consider, for instance, two successive tactile stimuli at two different spots on your arm. While the nature of the stimuli, the mechanical force, remains "unknown" to the neural system, as it gets transduced into neural activity, the spatiotemporal proportion of the stimuli can indeed be transferred. Both the spatial separation of the stimuli on the sensory surface, the skin, can be transferred into the neural system in terms of the spatial separation of two differently activated neural fibers as well as the temporal sequence of the stimuli in terms of the temporal sequence of the thus elicited neural spikes. And the same is true for other types of sensory stimulation. For instance, two successive visual stimuli at two different locations of the retina. Here again the nature of the stimuli, electromagnetic interaction in terms of photons, remains unacquainted to the neural system, while we may indeed consider the neural system to be "directly acquainted" with the spatiotemporal proportions of the stimuli. Hence, whatever the nature of the external stimuli, the spatiotemporal proportions of the stimuli can (but need not) be conveyed over the 3 N boundary.[6] We may think of this as the "spatiotemporal grounding of N in W via perception." Consider one more case. The various types of mechanoreceptors in the skin respond to mechanical stimulation such as pressure, stretching, and vibration. Clearly, the receptor signals cannot encode the nature of the external mechanical stimulation, but they encode the spatial change in the mechanoreceptor itself. No doubt, it is of utmost importance for our spatiotemporal grounding that mechanoreceptors can be found almost everywhere in the body, playing a role not only for exteroception but for interoception (somatosensation, proprioception, and viscероception; see the section "The neural/non-neural interface") as well.

[4] Note that because of the design of the example, in this case we know more than just cardinality, since we know that it is an order relation. So we know, for instance, that it is transitive.

[5] The basic idea of the Newman problem reappears as Putnam's (1976) well-known model-theoretic argument against metaphysical realism. See Bas van Fraassen's insightful discussion of the connection between Newman and Putnam, but also the relationship to the writings of Hermann Weyl, Rudolf Carnap, and, of course, Russell (van Fraassen 2008, Part 3).

[6] Here is a very basic consideration from which to look at things: brains and neurons exist in space and time. It is this simple fact that implies that neurons, like everything else in the world, also have spatiotemporal properties. Therefore, it is the nature of spatiotemporal relations, and only their nature, that can pass the 3N boundary.

Sensory intersections

Indeed, the previous consideration can be reinforced. As we have seen, unlike any other properties or relations, spatiotemporal relations may cross the 3N boundary. This, however, is only the first step to establish a connection between N and W. And luckily so. For it seems that even if a connection has been established between spatiotemporal relations in N and W, no such connection exists for the wealth of non-spatiotemporal signals as, for instance, the electromagnetic signals in the form of photons that reach the retinal receptors or the chemical signals the reach the taste receptors of the tongue.

To reinforce our point, we must remind ourselves of the fact that the different sensory organs or channels overlap or intersect in important ways. One aspect of this is that sensory organs are typically sensitive to stimuli of different nature. Retinal receptors are, for instance, also receptive for (strong) mechanical stimulation (leading to cloud-like visual impressions). Very bright light may be painful and, in audition, loudness or high pitch may be painful as well.

Another aspect is that many stimuli evoke responses in different sensory channels. I may see a black surface exposed to sunlight, and at the same time I can feel it as warm or even hot.[7] Likewise, I can feel the sugar cube on my tongue and it tastes sweet. And I can hear a deep bass tone and feel it in my stomach. Taken together, these sensory intersections (not to speak of the extreme case of synesthesia) serve to support each other in calibrating our various senses and relating them onto each other in a systematic and orchestrated fashion. Of course, the importance of multi- or crossmodal perception is long known for the issues of internal integration, binding, and unity (cf. O'Callaghan 2012; Bayne and Spence 2015). But here the point is that it is also important for matters of grounding. None of our sensory systems works in isolation. Rather, the whole neural system, the system beyond the 3N boundary, operates as an integrated whole. By means of the various sensory intersections, the "spatiotemporal grounding of N in W" infiltrates and pervades the entire neural system. It thereby leads to a grounding of the entire system. This is the ultimate reason why we can legitimately assume that neural mental representations of the external world are possible.

Neurophenomenal structuralism

This second part of the paper serves to introduce the relevant concepts and definitions of neurophenomenal structuralism. After presenting the central technical notions in the section "Structure, homomorphisms, and structural individuation" (much in line with FKL 3.1–3.2), the concepts of phenomenal content and character will be discussed (see the section "Phenomenality, content, and characters"). This should elucidate the first assumption of neurophenomenal structuralism. Next, I introduce quality spaces (see the section "Quality spaces, Q-structure, and psychophysics"). Thereafter, the concepts of neural activation spaces and neural maps will be assessed (see the section "N-structure: neural activation spaces and neural maps"). This elucidates the second assumption of neurophenomenal structuralism. The second part concludes with an evaluation of the available empirical evidence for structuralism and the prospects for the neuroscience of consciousness (see the section "Empirical evidence and consciousness studies").

Structure, homomorphisms, and structural individuation

In neurophenomenal structuralism the notion of structure is clearly central. Given the various state spaces of the domains W, Q, and N, the focus of neurophenomenal structuralism is on the structure of such spaces. This has already been motivated in the first part for the N-domain: neural systems work on a purely relational basis—the relational properties of neuronal activities. This idea can also be captured by saying that it is only the structure of N that counts. The idea can be elucidated by the definition of a so-called "relational[8] structure" as a set of relata with a set of relations imposed on them. More precisely, a structure is a tuple $\Sigma = \langle \alpha, R(\alpha) \rangle$ consisting of a carrier set or domain $R = \{R_1, R_2, \ldots, R_k\}$ of k elements or entities α_i equipped with a set of n-ary relations $R(\alpha)$. Therefore, to capture a domain structurally is to individuate the entities or relata only via the relations in which they stand. Given two domains, we can map entities in one domain onto entities in the other. Some mappings will also preserve the structures in a domain. The structures are then "structurally similar." In general, two structures A and B are structurally similar if the corresponding relations in A and B have the same number of arguments. Paradigmatic cases of structural similarity are the relationships between maps, pictures, and sculptures and what they represent or refer to. Structural similarity is mostly a second-order rather than a first-order similarity (cf. Shepard and Chipman 1970; O'Brien and Opie 2004). The latter consists of shared properties in both the source and the target domain. For example, colored chips are used by interior designers to select the intended color for painting a room. In the case of the former, the second-order similarity, it is only the relations between the relata that are shared. In a bar chart, for instance, rectangular bars or columns are used with heights proportional to the data that they represent; and in a weather map the spacing of isobars corresponds to pressure gradients in the atmosphere.[9] Structure-preserving mappings are mathematically known as "homomorphisms." More precisely, two relational structures $\Sigma_a = \langle a, R(a) \rangle$ and $\Sigma_b = \langle b, R(b) \rangle$ are of the same type, if for every relation $R_j \in R$, there is a corresponding relation $S_j \in S$ with the same arity. Given two such structures, a mapping h: $a \rightarrow b$ is a homomorphic mapping or homomorphism from Σ_a to Σ_b, if for every relation R_j and any elements $a_i \in a$ the following implication holds: $R_j(a_1, a_2, \ldots a_k) \Rightarrow S_j(h(a_1), h(a_2), \ldots h(a_k))$. Bijective homomorphisms are called "isomorphisms." If two structures are

[7] This point has a nice structural resemblance to a recent debate about the alleged underdetermination of ray optics connected to Newton versus Goethe. More precisely, Müller (2016) has launched an attack on Newtonian optics based on an alternative Goethe-style account which is intended as the direct opposite of Newton in assuming "darkness rays" as (i) opposed to light rays, (ii) consisting of complementary colors, and (iii) differing in degrees of refrangibility belonging to particular complementary colors. This modern Goethe account draws on a fascinating experimental realization of the "Goethe spectrum" (the spectrum of a bar rather than a slit) by means of a rigorous, full inversion of the optical space of the entire experimental setup of Newton's infamous *experimentum crucis* (Rang et al. 2017). Müller sees the Newton-Goethe case as a strong and genuine case of theory underdetermination. At best, however, the case should be understood as a hypothetical case of transient underdetermination, as it is decidedly restricted to the rather limited physical domain of classical ray optics (Lyre 2018b). By no means can the account be embedded into the bigger physical context. This can already be seen from simple everyday observations, in fact the analog of sensory intersections: Physical bodies get heated by sun light, and "darkness rays" are not connected with any heat or energy transport whatsoever. The Goethe account is therefore in severe conflict with thermodynamic energy conservation and does hardly connect to the later developments of electrodynamics.

[8] Structured sets may come equipped with either functions or relations or both. If a structure contains no relations, it is an algebraic structure; if it does not contain any functions, it is a relational structure (or system of relations). In this paper, the term structure is used in the sense of relational structure.

[9] There are also mixed forms of first- and second-order similarities. A true-to-scale road map, for example, represents spatial distance relations in nature by corresponding spatial distance relations on the map. However, the map is made of a different material than the objects it represents.

isomorphic, then they are structurally indistinguishable. In other words, any structure is defined only up to isomorphy.

Again, to capture a domain structurally is to individuate the relata by the relations in which they stand. A most important aspect of "structural individuation" is that Leibnizian individuality and intrinsicality must be given up (cf. Lyre 2018a). Graph theory provides a useful tool to illustrate this. Consider the example of an unlabeled and undirected graph with three nodes and three edges. Neither the nodes (relata) nor the edges (relations) are specified any further. This is analogous to the case of an "abstract" structure as considered in the section "The Newman problem of the brain and a proposal for its solution." As we have seen, in order to avoid the Newman problem we must specify the nature of the relations. Suppose that in our example the relation is that of equidistance, then the graph still applies to myriads of entities (physical objects, people, whatever) that realize the "spatial pattern" of an equilateral triangle. Hence, the structure represented by such a symmetric and specifically labeled graph is still multiply realizable. This is the idea of structural individuation: the nature of the relations is given, but the relata are determined up to isomorphy only. Of course there are cases where a structural individuation suffices to individuate the relata: namely, if the graph has no symmetries. Consider, for instance, a toy universe of just three otherwise unspecified relata. Suppose that they instantiate the distances of 3, 4, and 5 m among each other. They are then, so to speak, "maximally structurally individuated." Still, the structural individuation does not amount to the attribution of intrinsic properties to the relata.

In the following, it will be important that the nature of the relations of the Q- and N-structures is, at least partially, specified. Of course, this is not to address the trivial fact that in the case of N-structures we are dealing with relations (and relata) of neural nature, but for what such structures stand for, i.e. what they represent. Some such relations of the Q- and N-structures stand for spatiotemporal relations in W, since they can transcend the 3N boundary (as we saw in the "The Newman problem of the brain and a proposal for its solution" section). They then serve to ground structural mental representations (compare the section "Structural representation"). Against the background of what has been said above, we can now repeat the two main assumptions of neurophenomenal structuralism:

1. Any phenomenal experience is fully individuated by its place in a Q-structure. Phenomenal content and character are relational, not intrinsic properties of sentient subjects.
2. Q-structure is mirrored in N-structure. The mirroring has to be understood in terms of structural similarity, more precisely, in terms of a surjective homomorphism from N to Q.

The first assumption will mainly be treated in the section "Quality spaces, Q-structure, and psychophysics", while the second assumption will be treated in the section "N-structure: neural activation spaces and neural maps". But first, we need to make a few conceptual clarifications in the next section.

Phenomenality, content, and characters

Neurophenomenal structuralism is a special view about the phenomenal and the particular connection between the phenomenal and the neural. Of course, we have to clarify what is meant by "the phenomenal" and "the neural." In this section I consider "the phenomenal" first.

Let us start with terminology. Without loss of generality, we may simply assume the external world W to consist of objects that are in object states. Such object states are assigned object properties. So much already about the object side. On the subject side, we have creatures equipped with sensory or perceptual systems that can be in sensory or perceptual states. Crucially, such states are assigned qualitative or phenomenal properties, as such states can be experienced. The qualitative or phenomenal properties are therefore properties of the sense impressions of sentient beings.

A remark about this terminology. I use the terms "qualitative" and "phenomenal" as essentially synonymous. For certain purposes, however, it might be useful to distinguish between them. Clark (2000, 1.1) distinguishes between two types of properties as occurring in sensory experience. He construes qualitative properties in much the same sense as we do here, but considers phenomenal properties as properties attributed to objects in the world, hence, as such objects appear to us. On this count, qualitative properties are still properties of internal, sensory states, while phenomenal properties are properties of external object states. This usage of the phenomenal is reminiscent of Kant's usage of "phenomenon" (as opposed to "noumenon") as well as of the phenomenological tradition. I think that it is a useful distinction for certain purposes,[10] but it is not needed for our purposes here. Hence, I shall use phenomenal and qualitative as essentially synonymous.[11] Now, sense impressions have "content." They convey information about the external world (as we have seen in the first part) and serve as starting points for mental representations of the world. Hence, the qualitative or phenomenal properties capture the content of sensory or perceptual states. What kind of content is conveyed by perception? The crucial assumption of neurophenomenal structuralism is that "the specific content of a perceptual state consists in the structural facts encoded in the totality of all discriminations that a sensory system is able to perform." This is in accordance with our considerations in the first part where we saw that neural systems are based on change detection and relational coding. I will explain this in more detail below in this section, but we shall later also see how this speaks in favor of a modest representationalist position that takes only the second-order structural aspects of perceptions as potentially content-bearing. Indeed, being content-bearing, perceptual experiences must have accuracy conditions. They are not fulfilled in cases like illusions and hallucinations. Such experiences are inaccurate. To spell out the accuracy conditions of perceptual experiences would be the task of an account of perceptual content in terms of "structural representations." This is not the task of the present paper, but the topic of structural representations will be addressed in the section "Structural representation."

On the one hand, sense impressions have content, and on the other hand, sense impressions of sentient beings are frequently said to have a qualitative or phenomenal "character." Generally, to say that a mental state has a phenomenal character is to say that there is something it is like to be in that state. Qualitative or phenomenal properties must then capture such what-it's-like characters (also called "qualia" by many with a somewhat mysterious ring). Character and content are typically considered to be distinct. Neurophenomenal structuralism aims to connect both

[10] From this perspective "neurophenomenal structuralism" is a misnomer. But for lack of a better term I stick with it.

[11] Clark also proposes a feature-placing view according to which perceptual representation "proceeds by picking out place-times and characterizing qualities that appear at that place-times" (Clark 2000: 75), i.e. to "project" qualities to physical space in time. The account also suits the notion of topographic feature maps (see the section "N-structure: neural activation spaces and neural maps"). I'm very sympathetic to this view, but neurophenomenal structuralism is not committed to it, so I do not follow this line of reasoning here.

concepts. As the content of sensory states is rooted in structural aspects of perception, so does character. More precisely, the specificity of phenomenal character lies in the specific content it has. Since the specificity of perceptual content consists in the structural facts encoded in the totality of all possible sensory discriminations, the same holds for the specificity of phenomenal character.

There is, however, more to perceptual states than their specificity. In general, what makes mental states phenomenally conscious states, is, in addition, also the result of the neural mechanisms that underlie consciousness in general. As a working definition, we may say that phenomenal character is "experienced content or phenomenally conscious content." And neurophenomenal structuralism is an account that addresses the specific content of phenomenal experiences. It does not, however, address generic consciousness. In other words, it does not tell us when a mental state is conscious in general, but rather what the specific content of a conscious state is. As we ultimately strive to connect the phenomenal with the neural, we may also express this by using the terminology of neural correlates of consciousness (NCC). What has just been said is then in very good agreement with the claims of Marvan and Polák (2020) "that the perceptual NCC be divided into two constitutive subnotions. The first subnotion covers the content-specific side of the perceptual NCC… [T]he second subnotion [is] that of the neural process or processes making the perceptual contents… conscious." The authors call the first subnotion the neural correlate of content (NCc) and the second the general neural correlate of consciousness (gNCC). Hence, in the parlance of Marvan & Polák, neurophenomenal structuralism is about NCc rather than gNCC. And the above working definition of phenomenal character as experienced content is in full accordance with their view that the neural correlate of conscious content (NCcc, or, in our phrasing, the neural correlate of phenomenal character) "is a composite" of NCc and gNCC. Under this count, what it is like to see red? It is to be a specific location in the totality of all possible color experiences and simultaneously to be the target of a general mechanism of consciousness. Neurophenomenal structuralism is, in principle, compatible with various general approaches (e.g. Global Neural Workspace and Higher-Order-Theories, as will also be pointed out in the conclusion). What is crucial for our present considerations is that character is as much relational as is content. Why? Because character is a composite of specific content and a general mechanism. But the latter is at best system-intrinsic, not state-intrinsic (while the former is relational). Therefore, character is as much relational as content, and most of what follows will apply equally to content as to character. A final remark about our terminology: In light of the above our first assumption that phenomenal experience is "fully individuated" by its place in the totality of all possible experiences may be seen as a kind of shorthand for "fully individuated in its specificity." Indeed, what fully constitutes phenomenal experience (in the sense of a complete ontological grounding) is a composite of NCc and gNCC. However, since specific individuation depends only on the NCc part, we can omit the caveat "in its specificity" without loss.

Quality spaces, Q-structure, and psychophysics

We strive for the thesis that any specific qualitative character essentially consists in the structural facts encoded in the totality of all discriminations that a sensory system is able to perform. Which sensory system? The sensory system that brings about the sense impression of having the specific character in question. So, for instance, the qualitative or phenomenal character of a particular color impression consists in the structural facts encoded in the systematic totality of all possible color impressions. In the case of color, this totality is called the "color space." In general, it is called a "quality space or Q-space" (cf. Clark 1993 and Rosenthal 2010, 2015).[12] Quality spaces can in principle be constructed for any modalities (see Shepard 1982; Janata 2007; Collins *et al.* 2014 for a discussion of Q-spaces for the auditory system; Koulakov *et al.* 2011; Young *et al.* 2014 for the case of olfaction). Color is the by far most studied example, and in this paper I will restrict myself to it. We have thus encountered the first assumption of neurophenomenal structuralism: any (specific) phenomenal experience is fully individuated by its place in the structure of a Q-space (or Q-structure, for short). This elevates qualitative character to a relational rather than an intrinsic affair—quite contrary to the traditional view. Qualitative experience, as traditionally conceived, is considered an intrinsic property of the experiencing subject. Neurophenomenal structuralism strictly differs from this view by claiming the opposite: the qualitative characters of sensory experiences are relational rather than intrinsic properties of the sense impressions of sentient beings. Similar relational or structural views have been articulated by several authors (Raffman, 2015; Isaac 2014; Papineau 2015; Gert 2017a, 2017b), but most explicitly by Clark (2000, 18):

> So 'orange' could at best be defined as something like 'a reddish yellow, equally similar to red as to yellow, complementary to turquoise, more similar to red or to yellow than it is to turquoise', and so on. Each of the other qualitative terms would receive an equally enigmatic treatment. […] A consequence of this account is that qualitative character is a relational affair. Qualitative properties seem to be intrinsic properties, but they are not.

Indeed, the merit of structuralism lies precisely in the weakness of the opposite view: intrinsicalism. If we think of color qualia as strictly intrinsic—and thereby unrelated—then how should we ever account for color similarities? Hence, structuralism turns the traditional view upside down and starts immediately with the assumption that qualitative experiences are relational and can be unfolded in a Q-space structure. Although neither a clear proof for structuralism nor a knockout argument against intrinsicalism will be put forward here, the higher conceptual and empirical plausibility of the structuralist view should become apparent as the paper progresses.

Further motives that speak for structuralism can be found in the writings of the following authors: Papineau (2015) reminds us that even though humans are said to be capable of distinguishing millions of colors, they are not able to distinguish them in an absolute fashion. This already hints at the predominance of color relations. Alistair Isaac takes a measurement perspective on perception and argues "that the structure of our possible experiences corresponds to the structure of possible ways the world can be," but that "this structural correspondence … is calibrated differently across different contexts" (Isaac 2014, 508). This can be seen, for instance, in the well-known case where both a cooled and a warmed hand "are thrust into a lukewarm bucket of water" with the result that "the cool hand will sense the water as warm, while the warm hand will sense the water as cool" (Isaac 2014, 482).

[12] See also Raffman (2015) and the related but more general notion of a *conceptual space* by Gärdenfors (2000). See, furthermore, Decock (2006) for a physicalist interpretation of "phenomenal spaces" that comes close to the reductive physicalism advocated in the section "Reductive physicalism."

Crucially, from a structuralist standpoint, both sensations may be veridical as they correctly represent the difference between the antecedent temperature of each hand and the temperature of the bucket. Rather aptly, Gert (2017a, 6) speaks of brutely relational processes:

> Brutely relational processes yield relational information directly, without inferring it from non-relational information about the relata. As it turns out, human color vision also makes heavy use of many brutely relational processes. For example, the sort of opponent processing that is performed by ganglion cells, the receptive fields of which include both a region that is inhibited and a region that is excited by illumination, can detect certain edge contrasts without detecting absolute information about the intensity of the light reflecting from either side of the edge. This allows for the detection of such brutely relational facts as the following: that the surface on one side of a boundary is slightly lighter than the surface on the other side. And similar edge contrasts occur further along in visual processing.

The crucial point is that we should not consider Q-structures or Q-spaces as representing absolute values, but rather as a means to represent relations. They encode and represent the totality of possible sensory discriminations. Clark (2000, 251–252) has spelled out a further remarkable consequence of this:

> …in sensory systems the root relation from which order derives is not similarity, but difference. Instead of qualitative identity as the primitive term, the neural reality would be better reflected by using discriminability: the detection of differences. In sensory terms, similarity is naught but a failure to discriminate. Discrimination carries the load. As its dual, similarity gets a free ride. […] We do not identify what property is strictly identical among all the blue things, or that wherein all the blue things are the same; but instead that which distinguishes blue from all the other colours. 'Blue' marks that set of distinctions.

So, according to Clark—and I happily follow—discriminability and similarity (more properly: dissimilarity) are two sides of the same coin, with discriminability as root relation. This echos our considerations in the section "Change detection and relational coding" on the primacy of change detection and relational coding. In the following, similarity, or more precisely, similarity ratings, will be of interest.

To understand the idea of Q-structure even better, we should briefly turn to psychophysics. Psychophysics investigates the relationship between W and Q, between physical stimuli and sensations, and it aims to quantitatively determine or measure the impact that systematic variations of stimuli have on the sensations they produce (cf. Gescheider 1997; Lawless 2013). The field is rooted in the works of Ernst Heinrich Weber and Gustav Theodor Fechner in the 19th century. Weber's law, as heralded by Fechner, tells us that the smallest perceivable stimulus change dS is proportional to the initial stimulus intensity S, formally $dS \sim S$. The quantity dS denotes a "just noticable difference" (JND). For example, the JND for weights in humans is about 2% (within reasonable limits), i.e. a weight of 50 g can only be distinguished from 51 g, whereas a weight of 100 g can only be distinguished from 102 g.

In principle, JNDs provide elements or units for discriminability. It is, however, not enough to have pairwise discriminations, we want to somehow order our experiences. This is what similarity ratings do for us. The early color wheels or color circles that date back to the 17th century (and especially to the work of Newton) provide a standard example. It is no coincidence that the similarity order expressed in a color circle mirrors the ordering of colors in natural phenomena, notably in the prism spectrum, which is literally before everyone's eyes in the form of the rainbow (I come back to this in the section "Structural representation"). The same is true for the ordering of pitch following a frequency-related order of sounds. Remarkably, however, the perceptual system compactifies the experienced color order by transforming the linear order into a circle. It thereby changes the topology so that the two ends of the visible spectrum, violet and red, become neighbors in the color wheel with shades of purple in between.[13] Considered as a mathematical space, this ordering scheme is still one-dimensional but curved in itself. To account for discriminations and similarity ratings related to colorfulness (more technically, saturation or chromatic intensity), a second dimension must be added, thereby transforming the color wheel into a two-dimensional color disc. Finally, to account for all possible discriminations, a third dimension related to lightness must be added, thus creating the full color space spanned by the three dimensions hue, saturation, and lightness.[14] With JND's as units and similarity ratings, it is in principle possible to construct and to map out the color space. One result is the well-known CIELAB (more precisely CIE L*a*b*) color space that has also proven useful for technological purposes.[15] The method of similarity ratings combined with JND's works analogously for other modalities. Hence, the totality of all possible sensory experiences for some particular modality can be represented by a particular Q-space or quality space as already introduced in the previous section.

To sum up (and, at the same time, look ahead): "The specific content of sensory states consists in the facts encoded in a Q-structure understood as a second order structural representation mirroring a corresponding vehicle N-structure" (the second part of this statement refers to the section "N-structure: neural activation spaces and neural maps"). Since phenomenal character is experienced content, the specific what-it's-like character of our sensory experiences consists likewise in the facts encoded in a Q-structure understood as the totality of all possible sensory discriminations. This is a strong and revisionary metaphysical claim of neurophenomenal structuralism, but we shall see that it can also be linked to methodology. In consciousness research, a common and widely used method is to employ (subjective or objective) threshold measurements to determine conscious or unconscious stimuli. However, rather than asking whether a particular stimulus is conscious or not, it seems far more appropriate, from the perspective of neurophenomenal structuralism, to ask how a given stimulus relates to other stimuli. This shall be taken up in the section "Empirical evidence and consciousness studies."

[13] A first indication that colors are Q- and not W-properties, since there is no such thing as monochromatic "purple light." More precisely, there is no such thing as a single electromagnetic wavelength that evokes purple experiences.

[14] A note on the terminology: The dimensions of hue, saturation and lightness span a three-dimensional space, which is often referred to as "color space." The totality of all possible color experiences is of course only a region or a finite subspace of this total space. It has a highly complex and asymmetrical shape and is known as the "color solid" or the "color spindle." Very often, however, the color solid/spindle is also referred to as the color space. This latter use will also be the typical use of the term color space in this paper. And this generalizes to the usage of the terms Q-space and quality space.

[15] There are other color order systems or color spaces available (e.g. the Munsell color system, uniform color space, various CIE spaces; cf. Kuehni and Schwarz 2008). They mainly differ in terms of their experimental origin and the mathematical transformations of the various appearance parameters.

N-structure: neural activation spaces and neural maps

Having explored the structure of the Q-domain, we must now turn to the corresponding N-domain. This brings us to computational neuroscience and the computational modeling of neural systems in terms of neural networks. Indeed, the most straightforward way to think of N-spaces is in terms of the activation spaces of neural networks. Given a network of n neurons, the activity of each neuron can be represented on a corresponding axis, spanning an *n*-dimensional activation space. A distributed neural representation (i.e. the distributed neural activity of large populations of neurons) can formally be seen as a point in this state space (or a vector with the respective single activations as components). Such neural representations are generally understood as providing the neural vehicles of the content of neural states. According to Churchland (1989, 2001), this leads to a state-space semantics. This means that relationships in content are encoded in the distance relationships of neuronal population states in the activation space. In other words, the metric of the state space has a semantic interpretation.

Another type of neural representation is given by means of neural maps. This type of neural representation is of special interest for our purposes. Neural maps are ubiquitous in the brain. The best known class are cortical maps (cf. Bednar and Wilson 2015). We may distinguish two types of neural maps:

- Type-1: Ordered projections from a receptive surface onto some brain area, where the ordering benefits from using the surface topology as a principle of organization.
- Type-2: Ordered projections from (a class of) receptor neurons onto some brain area, where the ordering benefits from using the feature topology as a principle of organization.

Well-known examples of the first type are the tonotopy in the auditory system, the retinotopy in the visual system (in both LGN and the primary visual cortex), or the somatotopy in the somatosensory cortex. Examples of the second type are the "structural neural correlates" (in the parlance of FKL) or vehicles of Q-spaces, as, for instance, the color area V4.

It is common to neural feature maps that principles of self-organization may be exploited to develop computational models of such mappings. These models are known as self-organizing maps (SOMs). They are neural nets trained by unsupervised learning. The first SOMs were models for the development of orientation columns (von der Malsburg 1973) and ocular dominance columns (Willshaw and von der Malsburg 1976). Kohonen (1982, 2001) presented self-organized topographic feature maps (Kohonen maps). The key idea is that neuron weights should map input "feature spaces" (i.e. their data structures) by arranging themselves according to the topology in the feature space.[16] Hence, Kohonen maps serve to implement the rather general idea that neighborhood in feature space can be encoded by means of neighborhood in the neural wiring. Likewise, they reduce dimensionality by converting the nonlinear statistical relationships between high-dimensional data into the geometric relationships of a low-dimensional neural wiring, typically a two-dimensional neural map. Consider the color similarity structure as mirrored in a neural map (say color area V4). On the sensory surface, the retina, we find neural receptors that are sensitive to certain ranges of electromagnetic wavelengths (S, M, and L cone cells). Rather than using the spatial neighborhood on the retina, as exploited for the retinotopic projections from the retina to LGN or V1 (type-1 neural maps), the possible discriminations of cone cells, their change detections, serve to develop a neural map (type-2). Roughly simplified, the story is as follows: The three cone cells span a three-dimensional feature space. The wiring of the cones amounts for the color opponency mechanism, effectively a mathematical transformation of the input.[17] The three-dimensional feature space is now given by the three axes blue-yellow, red-green, and black-white. We may think of the further wiring and projection to the cortex in terms of an SOM. The result is a neural map that mirrors the structure of color space, since the SOM algorithm exploits feature topology as the principle of organization. The story can be generalized to neural maps mirroring arbitrary Q-spaces. Having clarified the core nature of Q- and N-spaces, we can now work out the full formulation of the second assumption of neurophenomenal structuralism. This assumption not only states that Q-structure is mirrored in N-structure but also postulates a special kind of structural mapping. The reasoning for this is straightforward. It is well known that neural systems are highly plastic and adaptive. Therefore, in order to account for neural plasticity and the notorious multiple realizability of Q-structure in N-structure, the structural similarity between N and Q must generally be assumed to be many-to-one.[18] Hence, the second assumption of neurophenomenal structuralism demands a "surjective homomorphism" from N to Q. Such surjective homomorphisms are also known as "epimorphisms."

Empirical evidence and consciousness studies

What remains to conclude the second part is to highlight the extent to which our assumptions are indeed in accordance with empirical evidence, and how structuralist ideas increasingly play a role in the recent neuroscience of consciousness. As already mentioned, neural maps are ubiquitous in the brain. Since SOMs exploit either spatial or feature topology as principles of organization, self-organized neural maps possess content that is encoded in the neural wiring of such maps. In other words (and in analogy to the state space semantics), the topology of neural SOMs has "a semantic interpretation." Both state space semantics and the semantic interpretation of neural maps can be viewed as variants of what has recently come to be known in cognitive and computational neuroscience as "representational geometry." The idea is that mental representations can generally be arranged in a representational space that is provided with a geometry based on the similarity of content. This is obviously close to the notions of conceptual spaces and qualia spaces

[16] According to Kohonen (2003, 1182), the general SOM algorithm can be set up in the following way: "In order that data-driven self-organization be most effective, the following two partial processes should always be implementable in as pure a form as possible.
(i) Find that cell in the network that matches best with the present input (in the sense of some criterion).
(ii) Modify this cell and its neighbors in the network to improve their matches with the present input."

[17] The color opponency mechanism makes simple scenarios of metamerism immediately obvious: the fact that different physical stimuli (different spectral power distributions) can lead to subjectively matching color perceptions. For instance, the additive mixture of a 620-nm red light and a 530-nm green light matches the color perception of a 580-nm yellow light (in healthytypical sub).

[18] Given two sets *A* and *B*, an injective mapping from *A* to *B* is a one-to-one mapping where we may also have elements of *B* without matching *A*-elements. A surjective mapping from *A* to *B* means that every *B*-element has at least one matching element of *A*, but maybe more than one. Hence, in general, a surjective mapping is a many-to-one mapping. A mapping that is both injective and surjective is called bijective. Therefore, a bijective mapping amounts to a one-to-one correspondence where each element of one set is paired with exactly one element of the other set.

(section "Quality spaces, Q-structure, and psychophysics"). In cognitive psychology, this can be traced back to works by Palmer (1978, 1999), Shepard (1968), Shepard and Chipman (1970), and Edelman (1998). A recent method to explore representational geometries is "representational similarity analysis" (RSA; cf. Edelman *et al.* 1998; Kriegeskorte 2008; Kriegeskorte and Kievit 2013). RSA is a multivariate pattern analysis (MVPA) technique (Haxby *et al.* 2014) that allows to compare two data sets of different origins by means of their so-called representational dissimilarity matrix, which is nothing but a measure of their second-order structural (dis-)similarity. Since the data sets may indeed be of entirely different origin, the method allows for comparisons between otherwise different fields, be it brain-activity data (such as fMRI, EEG, or cell recordings), behavioral data (such as reaction times or psychophysical similarity ratings), or data from computational models (such as deep neural networks). RSA is indeed tailor-made for the purposes of neurophenomenal structuralism,[19] and meanwhile there is ample evidence for RSA and related representational geometry methods illuminating the mutual dependencies in the triangle of neuroscience, cognitive science/psychology, and machine learning (cf. Diedrichsen and Kriegeskorte 2017 for an overview). An example that is much noticed and particularly interesting for our purposes is the study by Brouwer and Heeger (2009), who have used principal component analysis, another MVPA technique, to show that the color space similarity structure can be found in V4. Therefore, in FKL, we consider V4 as a candidate N-structure or structural neural correlate of the Q-structure of color experience. The general idea of using representational geometries for a "structuralist" understanding of consciousness is increasingly playing a role in the neuroscience of consciousness. Malach (2021) has recently argued that "structuralist ideas in which the content of a conscious experience is defined by its relationship to all other contents within an experiential category" speak in favor of a "localist perspective in which localized cortical regions each underlie the emergence of a unique category of conscious experience." Tsuchiya and Saigo (2021) and Tsuchiya *et al.* (2022) also strive for a relational definition of consciousness. For this purpose, they go as far as to make use of category theory, a part of mathematics that can be understood as a general theory of mathematical structures.[20] Tsuchiya *et al.* (2016) also try to build a bridge from the category-theoretic approach of consciousness to Integrated Information Theory (IIT).[21]

But apart from the prospects that such high-level abstractions offer for deciphering the secrets of consciousness, a final remark should be made about the "methodological" impact that structuralist ideas can and should have. Remember the first assumption of neurophenomenal structuralism, according to which any single experience is fully individuated by its place in the whole Q-structure. Taking this seriously actually affects the method and design of experiments in consciousness studies. The major and traditional way of making the distinction between unconscious and conscious stimuli experimentally accessible lies in the use of various types of thresholds together with the masking and priming of "individual stimuli." But, as, for instance, Tsuchiya *et al.* (2022) rightly point out, "given that any moment of conscious experience is never binary, this type of approach obviously oversimplifies the phenomenological character." Neurophenomenal structuralism, on the other hand, proposes measurement procedures that focus not on the properties of individual stimuli but on the relationships between different stimuli. Similarity ratings are, of course, a suitable example, as is the aforementioned RSA method. It is to be hoped that such "structuralist measurement procedures" will become established in consciousness research in the near future.

The structuralist's stance and its prospects

In this third part, relevant philosophical implications of neurophenomenal structuralism will be explored. I first like to clarify that it is not a version of structural realism (see the section "Structural realism"). Then, I argue why and in what ways neurophenomenal structuralism corresponds to reductive physicalism (see the section "Reductive physicalism"), and the concept of structural representation (see the section "Structural representation"). The first three sections serve to situate neurophenomenal structuralism in the larger philosophical landscape. In the remaining two sections I discuss the direct implications. We will see that neurophenomenal structuralism leads to neurophenomenal holism (see the section "Neurophenomenal holism") and that it serves to reject inverted qualia. A rigorous structuralist understanding of what-it's-like character can also be seen as a contribution to solving the hard problem (see the section "Qualia inversion and (finally) the hard problem").

[19] Of course, MVPA techniques such as RSA based on neuroimaging data give us no direct access on the neural subpopulations that serve as proper neural vehicles. According to Roskies (2021), they provide us with "proxy vehicles" and "provisional representations" only.

[20] In category theory, roughly speaking, one moves from the consideration of mathematical structures and their morphisms to the next higher level of abstraction and considers categories, i.e. classes of related mathematical structures together with their morphisms, and their relationships, functors, i.e. structure-preserving mappings between categories.

[21] This tempts me to say a few words about IIT. Unlike many other accounts of consciousness (e.g. global neural workspace, recurrent processing, or binding byneural synchrony),not specify a (class of) mechanism(s), but aims to highlight an information-theoretic measure as a parameter of consciousness (Oizumi *et al.* 2014; Tononi *et al.* 2016; Tononi and Koch 2016). In simplified terms, the integrated information of a system S measures the causal irreducibility of the system. It is an intrinsic quantity of S and positive precisely when the integrated information of all partitions of S is smaller than that of all of S. Integrated information thus captures what proponents of IIT call the "maximally irreducible cause–effect structure." Phenomenal experience is considered to be identical with this structure. In this crude sense, then, IIT seems to be a structuralist theory of phenomenality. The amount of integrated information determines the *quantity* of experience (but note that, as the number of possible partitions of a system grows exponentially, calculating this quantity for realistic neural systems and brains is practically impossible). IIT also introduces the notion of a *qualia space* (Balduzzi and Tononi 2009). For a system consisting of n elements with binary states this is a 2^n-dimensional state space. Any maximally irreducible cause–effect structure is characterized by its shape in this qualia space; and this is how IIT determines the *quality* of a phenomenal experience. At first glance, there seems to be a similarity between IIT and neurophenomenal structuralism. But there are actually some important differences. The first thing to note is that IIT sees itself as an account for both generic and specific consciousness, while neurophenomenal structuralism is concerned with the specific contents of consciousness only. Two other obvious differences: IIT proponents espouse a self-proclaimed panpsychism and allow for zombies. Space constraints prevent me from going into detail here, but this is where IIT's somewhat misleading terminology becomes a problem. As I see it, IIT is a misnomer, since it is not an information theory at all, but rather a theory about something like "causal irreducibility" (or "complexity" or "incompressibility" or something). To be sure, terms such as information, complexity, or entropy have long formed a conceptual jumble (see Ladyman *et al.* 2013 for an overview). It's perfectly alright to use information-theoretic measures to quantify things, and this is what IIT does (in terms of Shannon entropy and mutual information)—on the one hand side. On the other hand side, they insist that it is the physical cause–effect structure, the causal powers of the particular physical vehicle, that counts. This crucial feature simply cannot be adequately captured in information-theoretic language. Since physical cause–effect structures can in principle be inherent in all sorts of things, animate or inanimate, panpsychism follows. So does the claim that there can be zombies, since "any neural network with feedback circuits can be mapped onto a purely feed-forward network in such a manner that the latter approximates its input–output relationships" (Tononi and Koch 2016, 13). What is decisively altered by this type of mapping, according to IIT, is physical structure with inherent causal powers (misleadingly quantified as "integrated information"). To summarize: IIT is intrinsicalism about phenomenality, and thus the opposite of structuralism. Neurophenomenal structuralism does neither lead to panpsychism, nor does it allow for zombies (as should become clear from our reflections about metaphysical necessity and possibility in sections "Reductive physicalism" and "Qualia inversion and (finally) the hard problem").

Structural realism

It is advisable to clear up a possible misunderstanding right at the start. One might be inclined to ask whether neurophenomenal structuralism is a version of structural realism. The short answer is: no. In particular, neurophenomenal structuralism is no realism about the phenomenal.

What is structural realism (SR)? First and foremost, SR can be seen as a moderate version of scientific realism. Scientific realism, in turn, is the view that we should be realists about the theoretical entities posited by our most successful and mature scientific theories. SR then says that we should be committed to the structural rather than object-like content of our science, primarily the fundamental theories in physics (cf. French 2014 and Lyre 2010 for overviews). SR comes in different flavors. Some construe it as an epistemic doctrine (ESR), according to which the true nature of things in W, the external world scrutinized by science, is beyond our epistemic reach or capacities, while it is only structure that can epistemically be assessed. Others prefer an ontic version of SR (OSR), according to which there are no intrinsic natures of things anyway. Of these two alternatives, OSR is certainly the most widely adopted version in the context of fundamental physics (i.e. quantum theory, general relativity, and gauge theories). And here again the most widely adopted variant is a non-eliminative, moderate OSR, which states that there are relations and relata, but that there is nothing more to the relata than the relations in which they stand. This essentially means that we should repudiate the concepts of both intrinsicality and (strong Leibnizian) individuality in our fundamental metaphysics.

With this in mind, there seems to be a family resemblance to neurophenomenal structuralism, since our doctrine denies phenomenal intrinsicality and considers phenomenal experience as individuated by its place in a Q-structure (our first assumption). But crucially, and according to our second assumption, such a Q-structure is mirrored in N-structure. As I have argued in the section "N-structure: neural activation spaces and neural maps," this is plausibly to be understood as a vehicle thesis: phenomenal states can be attributed content with states of neural maps as the bearers of such content. Moreover, as I shall argue in the next section, we can run an exclusion argument according to which the Q-structure plays no causal role at all, but all the causal work is taken over by the N-structure. Therefore, neurophenomenal structuralism, as I advocate it here, is no realism about the phenomenal (be it structural or otherwise), and it is rather consistent with a reduction of the phenomenal to the neural, as we shall also see in the following section.

At this point, my view differs slightly from the (otherwise quite similar) view of Isaac (2014), who attempts to establish an ESR about color qualities. Crucially and unlike neurophenomenal structuralism, Isaac is interested in the relationship between the Q-domain and the W-domain, between perception and the world, and argues for ESR in this respect. Neurophenomenal structuralism, on the other hand, is a thesis about neither the Q–W relation nor the N–W relation, but only the Q–N relation. Indeed, neurophenomenal structuralism is initially quiet about the relationship to the external world, apart from the fundamental considerations in section "The Newman problem of the brain and a proposal for its solution" about the Newman problem and its proposed solution.

That said, we may nevertheless go a step further. While neurophenomenal structuralism is initially quiet about the relationship to the external world, it is nevertheless much in tune with a (modest and instrumental) representationalist picture. And sure enough, as already pointed out in the section "The Newman problem of the brain and a proposal for its solution," in order to represent, the representations must somehow refer to W. This leads us, in the section "Structural representation," to the concept of structural representation, but we must first address the issues of physicalism and reductionism.

Reductive physicalism

In general, to consider two domains as structurally similar doesn't render the two domains as being ontologically on a par. This should particularly be clear if the structural similarity is of second order (compare the section "Structure, homomorphisms, and structural individuation"). In case of the structural Q–N similarity, however, things are different. At first sight and unlike the N-domain, the Q-domain does not seem to be part of the physical world. But is that true? This raises the issue of reductionism. It is one of the issues that the FKL paper circumvents. We rather say that neurophenomenal structuralism "opens an attractive door for reductionism, but… [that] there may also be a non-reductive reading" (FKL, p. 10, footnote 18). Why non-reductive? One reason can be seen in the multiple realizability of Q in N as already indicated in the section "N-structure: neural activation spaces and neural maps," and another reason is the well-known fact that even under the reasonable requirement of Q–N supervenience, there is a loophole left for dualism.

Consider multiple realizability (MR) first. Applied to Q- and N-types, the classic MR argument against type-reductionism says that if Q-types are multiply realizable in N-types, then Q-types are not identical (cannot be reduced to) N-types. It is crucial for the argument that MR is here understood in the sense of drastically heterogeneous realizations. But what counts as a realization in the first place? Shapiro (2000)[22] has forcefully countered the MR argument with the following dilemma: If the N-types share many causally relevant properties, then they are not distinct realizations. If they have no or only a few causally relevant properties in common, then there are no or only a few laws that apply to all realizers. In this latter case, however, the Q-types are no genuine types, i.e. they do not pick out genuine property classes in the world at all. In my view, this is a strong and compelling objection against the classic MR argument. Pain, for instance, is no Q-type at all, as evidenced by the fact that there are no general "pain laws" (except for the analytical triviality that pain is painful). Shapiro's dilemma also indicates that realizers must not share many properties (let alone all), but only the causally relevant ones. Lyre (2009) has pointed out that realizers sometimes even share only relational (causally relevant) properties. The property of being a harmonic oscillator is seemingly MR (pendula, springs, oscillatory circuits, etc.) and also allows for the oscillator equation as a dynamical law. Yet the realizers are not drastically heterogeneous. They share causally relevant relational properties, since the dynamic variable in the oscillator equation describes relational quantities of change (of amplitudes, angles, concentration, etc.). Realizations of harmonic oscillators are therefore unproblematic reductionist cases of MR. It is now important to note that the same rationale applies to Q-types according to neurophenomenal structuralism. Phenomenal properties, Q-types, are fully individuated by their place in a Q-structure (for instance, orange experiences as individuated between yellow and red). Q-structures are indeed multiply realizable in N-structures. But Q-structures are also mirrored in N-structures, which precisely means that the realizing N-structures share the causally relevant relational properties. The MR argument against type-reductionism has no force

[22] See also Kim (1992) as similar in spirit.

in a structuralist setting of phenomenal properties. On the contrary, structurally construed Q-types can *a fortiori* be reduced to N-types.

The other alleged reason for "a non-reductive reading" of neurophenomenal structuralism touches on the fact that even Q–N supervenience leaves a loophole for dualism. At its core, the requirement of supervenience consists in a mere covariance claim: higher-level mental properties covary with lower-level physical properties. Curiously, many (if not most) versions of dualism are in accordance with this. It is the modal force of the supervenience claim that makes the difference. I claim that the second assumption of neurophenomenal structuralism should therefore be read as endowed not only with nomological but with metaphysical necessity. More precisely, proponents of neurophenomenal structuralism are well-advised to assume metaphysically necessary Q–N supervenience. In other words, it is metaphysically impossible (and inconceivable) that Q-properties do not supervene over N-properties. This rules out dualism and respects the "scientific spirit" of neurophenomenal structuralism.

There is, I admit, no argument for upholding Q–N supervenience with metaphysical necessity (and ruling out the metaphysical possibility of dualism) other than the request to adhere to the overall scientific, more properly physicalist, spirit of neurophenomenal structuralism. Remarkably, however, once we accept this type of supervenience, the door is open for a tightened argument in favor of reductive physicalism. And this is due to the combination of supervenience with our second assumption, the structural Q–N mirroring. This combination allows for what I like to call a tightened exclusion argument. Jaegwon Kim's well-known causal exclusion argument says that given causal closure of the physical and the supervenience of the mental on the physical, there cannot be any irreducible mental causes (Kim 1998). Here, causal closure means that every physical effect has a physical cause, if it has a cause at all. So unless we allow for (weird) overdetermination, the mental is causally idle and "excluded" by the causal work of the physical. Now, the Q–N supervenience "with metaphysical necessity"[23] says that there is no ontological gap between the Q- and N-domains—which is likewise a statement of physicalism. While such a supervenience already allows for a causal exclusion argument, the argument gets strengthened by the additional requirement of the mirroring of Q-structures in N-structures. As all Q-structure *ipso facto* includes the causal structure of Q, that causal structure is mirrored and hence preempted by the N-structure. Thus, the Q-structure plays no causal role at all; all causal work is taken over by the N-structure. All this, of course, fits very well with our previous reductionist considerations regarding MR.

In the section "Phenomenality, content, and characters," I have argued in favor of a view that considers perceptual states as content-bearing. Then, in the section "N-structure: neural activation spaces and neural maps," we have seen that the Q-domain is about the content, while the N-domain is about the content-bearing vehicles. Combined with the reductive physicalism just advocated, this results in a view that regard content as instrumental at best. This does not mean that we must abandon the notion of representation altogether. Undoubtedly, the notion of representation serves many practical purposes in the cognitive and computational neurosciences. And that is also the use I want to make of it here. But it also means that everything that counts as representational (in the realm of phenomenal experience) must ultimately be directly reconstructable in terms of the physical properties of the neural vehicles. This is the idea of a "vehicle theory of representation" (cf. O'Brien and Opie 1999, 2001 for such a vehicle theory in the realm of phenomenal experience), and it leads over to the next section.

Structural representation

It is natural to ask whether neurophenomenal structuralism is a version of the recent program of structural representation. The short answer is: yes, albeit of a very special sort.

So what is a structural representation (S-Rep)? Indeed, in the recent debate about the nature of mental representation in philosophy of mind and philosophy of cognitive science, the concept of an S-Rep underwent a renaissance.[24] A major reason for this renaissance is that the programs of naturalizing semantics by means of causal theories or teleosemantics are fraught with insurmountable problems.[25] S-Reps are based on structural similarity. Paradigmatic cases of structural similarity are maps, pictures, or sculptures. The criterion of structural similarity makes S-Reps a variant of the classical similarity conception of representation. This conception is subject to massive caveats and objections, most notably elaborated by Goodman (1976). The similarity relation is symmetrical, the representation relation is not: My passport picture represents me, but I do not represent my passport picture. One twin may be very similar to the other, but never represents him. And under very weak conditions of similarity, everything becomes similar to each other; conversely, similarity apparently depends on the observer or context.

To rebut these and further objections, proponents of structural representationalism postulate three conditions for S-Reps:

1. S-Reps are generally based on homomorphisms. The restriction to isomorphisms, i.e. bijective homomorphisms, albeit urged by many, is unnecessarily narrow. The mapping could violate either injectivity or surjectivity. It would then be non-symmetric, which is a general requirement for representations (Bartels 2006).
2. S-Reps must be successfully exploitable by the system (Isaac 2013; Shea 2014; Gładziejewski 2016; Gładziejewski and Miłkowski 2017). The background consideration is to separate the problem of content from the problem of use (Ramsey 2016).
3. S-Reps must be causally grounded in the world. Isaac (2013) calls for "causally grounded homomorphisms"; Shea (2014, 2018) considers a hybrid of teleosemantics and structural representation as part of his "Varitel" semantics.

The first condition is rather permissive: S-Reps based on structural similarity (homomorphisms) can be found almost everywhere. Therefore, the requirements of the second and third conditions serve as important restrictions. It is conceivable, for instance, that a subject's neural map is structurally similar to the spatial layout of, say, some lunar crater. However, without the subject ever having been on the moon (third condition of causal grounding) or ever having used the structural similarity (condition of exploitability), the neural map is not a representation of the said crater. Particular attention should be paid to the

[23] For the sake of brevity I will suppress this proviso in the following.

[24] Starting with (Cummins 1989, 1996; Swoyer 1991; O'Brien and Opie 2004 over Bartels 2006; van Fraassen 2008; Ramsey 2007 to Shagrir 2012; Isaac 2013; Shea 2014; Gładziejewski 2016; Gładziejewski and Miłkowski 2017).

[25] Most notably the problem of misrepresentation in the case of causal theories and the indeterminacy of function in the case of teleosemantics.

exploitability condition. It clearly shows that the program of structural representation is not a strict naturalization program, but rather a pragmatic or instrumentalist program. Indeed, the strict urge to naturalize semantics ultimately only arises for realists about mental content (as, for instance, in the case of intentional realism à la Fodor 1987).

Mutatis mutandis, the three conditions also apply to neurophenomenal structuralism. Here the first condition, which establishes a relationship to the external world W, deserves special care. To see this, it is instructive to distinguish three types of S-Reps concerning

1. spatial structure,
2. temporal structure, and
3. feature structure.

Paradigmatic cases of the first type are maps, as they draw on the static similarity with regard to the spatial structure of both the representation and the representandum (the represented object). An example of the second type would be the oculomotor system, as portrayed by Shagrir (2012). This neuronal system computes an integration function by converting input in terms of eye velocity into output in terms of eye position. The neural integration thereby mirrors the temporal structure of the dynamical relationship between eye velocity and eye position. In both type-1 and type-2 cases, the corresponding S-Reps mirror concrete spatiotemporal affairs of the external world W. In light of what I have argued for in part 1 (especially the section "The Newman problem of the brain and a proposal for its solution,"), both types of S-Reps rely on the spatiotemporal grounding of N in W via perception, since both types rely on difference coding in terms of spatiotemporal relations transferring the 3N boundary. Somewhat metaphorically speaking we can say that difference coding with spatiotemporal discriminability as root relation provides the primordial ground for a structuralist conception of the mind.

Type-3 cases of S-Reps differ from type-1 and type-2 since they do not mirror concrete spatiotemporal affairs in the external world. In fact, paradigmatic type-3 cases are the neural maps considered by neurophenomenal structuralism (see the section "N-structure: neural activation spaces and neural maps"). In such cases, abstract W-structure in terms of feature similarities is partially contained in the representing N-structure. Hence, neural maps as type-3 S-Reps are based on partial homomorphisms mirroring abstract (rather than concrete spatiotemporal) W-structure. This is the reason why the ordering of the rainbow, an abstract physical ordering in W that only becomes visible as a spatiotemporal affair because of the refraction of sun light in raindrops, is pertained in the human color space ordering. Or at least, it is partially pertained, as there is, for instance, no purple in the rainbow (blue and red are no neighbors in the prism spectrum; see the remarks about compactification in the section "Quality spaces, Q-structure, and psychophysics" and dimensionality reduction in the section "N-structure: neural activation spaces and neural maps"). And this can be generalized: neural feature maps mirror exploitable physical relationships, i.e. crucial elements of the W-structure that can be used for active behavior and the survival of the organism. Since the ordering is abstract, the W-N relation is of second order (cf. the section "Structure, homomorphisms, and structural individuation").

Rounding up the discussion and conclusions of the last three sections, the following line of argument results:

i. the contents of phenomenal experiences are encoded in Q-structures,
ii. Q-structures can be reduced to N-structures,
iii. neural maps are the N-structure vehicles of Q-structure phenomenal contents,
iv. neural maps are likewise type-3 cases of S-Reps,
v. S-Reps are in tune with instrumentalism about content and indicative of a vehicle theory of representation,
vi. S-Reps typically refer by means of second-order structural resemblance (between N-structure vehicles and W-structure),
vii. neural maps as type-3 cases of S-Reps do not directly represent spatiotemporal affairs, i.e. they only indirectly refer to the external world.

Neurophenomenal holism

Now that we have located neurophenomenal structuralism in the larger philosophical landscape, let us turn to its more direct implications. Structuralism about experience is more than just a construal of single experiences as relational. It is structuralism in the full sense (rather than mere relationalism), since the whole structure of a qualia space determines each internal relation. Any single experience is fully individuated by its place in the whole Q-structure. This, I claim, is equivalent to the requirement of a "neurophenomenal holism."

Think of the Q-structure as an elastic web. It encodes the totality of all experiences. Any single change, whether by addition or deletion of either nodes or links or both, has the potential to change large parts or even the entire web, so any single change results in adjustments to large parts or even the entire web. Therefore, and as a direct consequence, neurophenomenal structuralism leads to a particular form of holism. It leads to holism in much the same way as the Duhem–Quine thesis and the corresponding picture of science as a web of beliefs lead to confirmation holism (Quine 1951) or as functional role semantics leads to meaning holism (cf. Block 1998). The overall rationale is in both cases the same and truly structuralist: entities are what they are depending on the role they play in a net, web, or structure. In confirmation holism, this concerns the confirmation of single statements or hypotheses in the total web of scientific statements connected by logico-inferential roles. In meaning holism, this concerns the meaning of a word in a language connected by functional use roles or, even stronger, the intentional content of single beliefs and thoughts connected by inferential roles. However, while meaning holism concerning intentional content is widely discussed, it is seldom considered in the context of phenomenal content. This is precisely the idea of neurophenomenal holism.

One of the standard arguments against meaning holism is that meanings might become instable and cannot reliably be shared among subjects (Fodor and Lepore 1992). This argument can be countered with good reasons (Churchland 1993, 1998; Block 1998). But more importantly, this special type of argument has no force regarding neurophenomenal holism anyway. On the contrary, we should expect considerable variability in the Q-structure for members of a given species. Moreover, since the Q-structure is mirrored and ultimately grounded in the N-structure, the origin of the Q-variability lies in the physical changes of the underlying N-structure and the dependency of the neural development on individual or environmental circumstances. As we have seen in the Section N-structure: neural activation spaces and neural maps, the relevant N-structures, self-organized neural maps, have

a semantic interpretation in the spirit of a state-space semantics, which itself is a clear implementation of meaning holism (Churchland 2001). And so is the semantic interpretation of neural maps.

Again, we should expect considerable variability in the Q-structure. Consider once more the case of color experience (as also illustrated in FKL 3.3). Here, the Q-structure variation becomes evident in everyday life scenarios where people give deviating color judgments. An intriguing case is color blindness due to dichromacy, especially the common red-green color blindness. What can be said about the color experiences in such cases? Neurophenomenal holism predicts that subjects suffering from red-green blindness not only fail to distinguish red and green, but, compared to trichromats, will experience all colors differently, since their Q-structure differs significantly from the Q-structure of trichromats.

As a further example, consider audition. It is well known that human hearing diminishes with age. However, there are still pitches that an older person perceives as very high, only these are then pitches that the person experienced at best as high in her younger years, but not as *very* high. Yet people are typically not aware of this change over time. Because what changes are not single experiences, but the web or Q-structure of all possible experiences to each other. Older people are then occasionally made aware of this change by younger people, who report pitches that the older people can no longer perceive.

Finally, neurophenomenal holism is not limited to single modalities. In principle, it concerns the neural system as a whole, for two main reasons: first, because of the various sensory overlaps already highlighted in the section "Sensory intersections," and second, because of the extent to which no part of the downstream neural system can be strictly separated from all other parts. No sensory system operates in strict isolation, the various sensory systems can in principle influence each other. Typically, however, these mutual influences are weak. For this reason, the individual modalities can be considered independent for most practical purposes (with the notable exception, again, of synesthesia). Ultimately, it is an empirical question how strong the mutual influence and the accompanying holism really is.[26]

Qualia inversion and (finally) the hard problem

Qualia inversion scenarios provide a notorious class of thought experiments in philosophy of mind. Qualia inversion consists in the idea that it is possible to permute or invert the qualitative experiences in a systematic fashion (either intra-personally or inter-personally) despite the fact that the functional organization of the subject(s) remains invariant: Not only do the subjects involved in inversion scenarios exhibit the same overall behavior before and after the inversion, but there are no functional differences at all. Qualia enthusiasts consider such scenarios at least as metaphysically or, less common, nomologically possible. It is possible to have functional isomorphy and yet qualitative or phenomenal differences. In this way, qualia inversion serves as an argument against functionalism. Technically speaking, the idea of qualia inversion amounts to claiming that the N-domain allows for multiple "instantiations" in the Q-domain (rather than the other way round!). Therefore, the possibility of qualia inversion is likewise an argument against psycho-neural reduction.

What can be said about qualia inversion from the viewpoint of neurophenomenal structuralism? The most widely discussed cases of qualia inversion pertain color spectrum inversions. So, without loss of generality, I shall focus on color experiences. Philosophers have suggested various outlandish thought experiments as, for instance, Block's inverted earth (Block 1990). My considerations will be more humble, although. As we have seen in the section "Quality spaces, Q-structure, and psychophysics," color experiences are traditionally conceived as intrinsic properties of the subject. In contrast, neurophenomenal structuralism considers color experiences as individuated by their location in the structure of the Q-space for color, call it the color space C. They are strictly relational properties of the subject. And as we have also seen in section "Quality spaces, Q-structure, and psychophysics," color intrinsicalism has a hard time to explain the standard similarity ratings of color experiences: that, for instance, purple is closer to blue than to green and that orange lies in-between yellow and red. The C-structure is precisely an ordering and encoding of the totality of such color similarities.

Color inversion scenarios take advantage of the C-structure. More precisely, they take advantage of the possibility of internal symmetries of C. Technically speaking, they draw on the existence of a non-trivial automorphism group on C. The automorphism group of a geometrical object (as, for instance, the C-space) is the symmetry group of that object, i.e. the group of all transformations under which the object is invariant. The simplest and most frequently used case would be the full rotational symmetry of the non-distorted color circle as a highly simplified color space model. Let us call it C*. All possible color experiences must respect the structural relations encoded into C*. Because of the rotational symmetry it is, however, not possible to distinguish "within the C*-structure" the chain of yellow-orange-red from, say, blue-cyan-green. It might therefore seem then that structuralism is in tune with color inversion, but indeed the opposite is the case.

Three broad answers can be given with the above scenario as a starting point. I order them along the different possibility classes. First, and as many have pointed out (e.g. Hardin 1988, 1999; Clark 1993, 2000; Palmer 1999), unlike the simplified model C*, the real C-structure is highly asymmetric. Humans can discriminate, for instance, more shades of red and green than shades of yellow and blue. Also, red can become more saturated than yellow, i.e. red is less similar to white than yellow. So the first answer basically rules out the actual possibility of color inversion in human and, in fact, in all modalities and for all species with asymmetric Q-spaces. And clearly, it is an empirical question whether certain Q-spaces allow for certain (if only partial) symmetries or not. As things stand, no empirical evidence speaks in favor of this.

So color inversion is ruled out because of contingent facts. But can it be ruled out as a nomological possibility per se? The opponent might insist that it is in principle possible that C-structures allow for certain symmetries, if only partially. Indeed, I don't think that this can be ruled out on the basis of the laws of nature per se. But it can be ruled out on the basis of our considerations in sections "The Newman problem of the brain and a proposal for its solution" and "Sensory intersections"! Color perception, as any other perception, is grounded in the perception of spatiotemporal relations and the color system, as any sensory system, does not work in isolation. Color experiences are in a systematic fashion related to other experiences. For instance, dark-colored objects absorb heat better than light-colored objects. Color inversion, however, would be in conflict with this. This connects to

[26] Neurophenomenal holism might also be a reason why a definite answer to the famous Molyneux problem cannot be given so easily. The question whether someone born blind and familiar with the touch of a cube could recognize a cube upon seeing it later (after a spectacular cure from blindness) must probably be answered in a vague in-between way. Again in principle, no sensory system works in strict isolation. Typically, however, the mutual influences will be weak. To quantify this for all of the interdependencies between the various sensory modalities is an empirical task.

earlier points in this paper: the numerous sensory intersections within the overall internal neural system from the section "Sensory intersections" (including the footnote about the limits of the alleged Newton–Goethe underdetermination case) and the neurophenomenal holism in the section "Structural realism." All of these points in the same direction: qualia inversion of single modalities is incompatible with an overall holistic neural system bound together by sensory intersections and striving for internal consistence and coherence.

Finally, what is still left is the metaphysical possibility of qualia inversion. This view is not uncommon among strong qualia enthusiasts; it is the ultimate philosopher's fantasy. It means that we give up any law-like restrictions on the possible inversion scenarios: The qualia proponent may insist that subjects do indeed report orange as in-between yellow and red, but that they nevertheless experience some bizarre and utterly unsystematic combination of arbitrary colors, as the qualia identity is primitive and not settled by the causal or functional profile. One way to spell this out is to construe qualia as quiddistic properties (Chalmers 2012, Chap. 7.9). On such a conception, qualia, as quiddities, have a primitive property identity (the so-called primitive suchness) that is not settled by the property's causal or functional profile. As a metaphysical possibility, I fear, this cannot be ruled out. And it certainly cannot be captured by any structuralist conception. If this is the core difficulty about qualia, then the hard problem cannot be solved structurally. But maybe structuralism at least has an offer ready. Let us see.

Should we be worried by excessive metaphysics? David Lewis, in a widely recognized late paper on quidditism, recommends humility:

> …to reject quidditism is to accept identity of structurally indiscernible worlds – that is, worlds that differ just by a permutation or replacement of properties. […] It would be possible to combine my realism about possible worlds with anti-quidditism. I could simply insist that […] no property is ever instantiated in two different worlds. […] It could be for the sake of upholding identity of structurally indiscernible worlds, but I see no good reason for wanting to uphold that principle (Lewis 2009, 210–211).

Neurophenomenal structuralism is of course precisely a doctrine based on the principle of the identity of structurally indiscernible phenomenal worlds (not worlds per se—that would be a version of OSR, see the section "Structural realism"). Neurophenomenal structuralism upholds the identity of Q-space isomorphs. Why should one uphold such a doctrine? For one, it is of course far more in tune with our overall scientific understanding of the mind and brain (and I take it that even the strongest qualia enthusiasts have to admit this much). But I claim that it also offers a way to understand the what-it's-like character in terms of structural properties. The common similarity ratings (say, again, that orange is in-between yellow and red) tell us not only about phenomenal similarities, the specific what-it's-like character of an orange experience is, in its specificity (cf. see the section "Phenomenality, content, and characters), constituted by the position it has in the appropriate Q-space, the totality of all possible color experiences. On this count, what-it's-like character reduces to holistic phenomenal content (which in turn reduces to neural vehicle representations). This is at least what neurophenomenal structuralism has on offer to the notorious hard problem. And that offer fares far better than simply relocating what-it's-like character into mysterious intrinsic (or even quiddistic) qualia. To refer to what-it's-like character as primitive doesn't explain the hard problem at all, it just repeats its very statement in terms of mysterious and excessive metaphysics. Neurophenomenal structuralism is trying to do better.

Conclusion

We went a long way. Many issues have been addressed in this paper, and most of them deserve a much more detailed consideration. I am well aware of this, and the main excuse is that the paper was intended as a programmatic paper. I had to address a wide range of topics without being able to deal adequately with all of them. Therefore, the least I can do in this conclusion is to give a comprehensive summary of the main insights, claims and results that were developed in the course of the paper. In keeping with my programmatic purpose, but also to increase clarity, I take the liberty of presenting the summary in the form of a list of bullet points:

- The neural/non-neural (3N) boundary consists of an inward part of neural transduction and an outbound part of neuromotor transformation (see the section "The neural/non-neural interface").
- Neural sensory systems work by change detection, and only relational properties can cross the 3N boundary. Any of the neural coding schemata deliver difference coding solely based on the relational properties of the coding elements. These relational properties are the building blocks of the N-structure of the brain (see the section "Change detection and relational coding").
- The N-structure of the brain is threatened by the Newman problem, since both the intrinsic nature and the nature of the relations of the external W-causes perturbing N seem to remain hidden from the brain. The proposed solution is that spatiotemporal proportions of the external stimuli may cross the 3N-boundary, thereby ensuring the spatiotemporal grounding of N in W via perception (see the section "The Newman problem of the brain and a proposal for its solution").
- None of our sensory systems works in isolation. The numerous sensory intersections serve to support each other in calibrating and relating the various senses. In this way, the spatiotemporal grounding of N in W infiltrates and pervades the entire neural system (see the section "Sensory intersections").
- Structures are sets of relations defined over sets of relata. Structural similarity is second-order, where only the relations between the relata are shared. To capture a domain structurally is to individuate the relata by the relations in which they stand (see the section "Structure, homomorphisms, and structural individuation").
- Qualitative or phenomenal properties capture the specific content and character of sensory or perceptual states. Phenomenal content and character consist in the facts encoded in the structure of a Q-space (quality space) understood as the totality of all change detections that a sensory system is able to perform (see sections "Phenomenality, content, and characters" and Quality spaces, Q-structure, and psychophysics).
- This yields the first assumption of neurophenomenal structuralism: any phenomenal experience is fully individuated by its place in a Q-structure (see the section "Quality spaces, Q-structure, and psychophysics").
- Self-organized neural maps possess content encoded in their neural wiring. The topology of neural SOMs has a semantic interpretation. Neural maps are the structural neural

correlates or vehicles of phenomenality (see the section "N-structure: neural activation spaces and neural maps").
- This yields the second assumption of neurophenomenal structuralism: The Q-structure is mirrored in N-structure, where the mirroring is to be understood as a surjective homomorphism from N to Q (see the section "N-structure: neural activation spaces and neural maps").
- Neurophenomenal structuralism proposes measurement procedures in consciousness studies that focus on the relationships between different stimuli as, for instance, similarity ratings or representational geometry methods such as RSA (see the section "Empirical evidence and consciousness studies").
- Neurophenomenal structuralism is no realism, not even a structural realism, about the phenomenal (see the section "Structural realism").
- Proponents of neurophenomenal structuralism should assume a metaphysically necessary Q–N supervenience. Combined with a structural Q–N mirroring, a reductive physicalism follows (see the section "Reductive physicalism").
- The program of structural representation is in tune with an instrumentalism about mental content and indicative of a vehicle theory of representation. Neurophenomenal structuralism is a special version of this program, since neural maps are type-3 cases of S-Reps (concerning feature structure) that do not directly represent spatiotemporal affairs, but refer to the external world only indirectly (see the section "Structural representation").
- Neurophenomenal structuralism leads to neurophenomenal holism, since the whole Q-structure determines each internal relation, and any single experience is fully individuated by its place in this structure. Dichromats will therefore experience all colors differently, since their Q-structure differs significantly from the Q-structure of trichromats (see the section "Neurophenomenal holism").
- Neurophenomenal structuralism rules out the nomological possibility of qualia inversion. A rigorous structuralist understanding of what-it's-like character also offers a contribution for solving the hard problem (see the section "Qualia inversion and (finally) the hard problem").

Let me conclude with a follow-up on the last point. Neurophenomenal structuralism assumes phenomenal experience as individuated by its place in a Q-structure. As already pointed out in the introduction and more properly in the section "Phenomenality, content, and characters," this addresses the specific content of phenomenal experiences. Hence, neurophenomenal structuralism addresses specific rather than generic consciousness. Therefore, it can never fully solve the hard problem, but may contribute to its solution. To fully solve the hard problem, the general nature of consciousness must be uncovered. But here our program is open. Neurophenomenal structuralism is, in principle, compatible with various neuroscientific approaches such as Global Neural Workspace, Recurrent Processing, Higher-Order-Theories, Neural Synchrony and/or Critical Brain (modulo IIT, as seen in the section "Empirical evidence and consciousness studies"). My hunch is that elements of all the mentioned approaches will play a role, including the embodied and interoceptive nature of a self-model. But that remains to be seen. Neurophenomenal structuralism is ready to be included.

Acknowledgements

Many thanks to Sascha Fink, Alistair Isaac, Lukas Kob, and Marlo Paßler for valuable remarks and discussions about earlier versions of the paper. Special thanks also for very helpful comments by two anonymous reviewers of this journal.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 337619223/RTG2386.

Conflict of interest statement

None declared.

References

Balduzzi D, Tononi G. Qualia: the geometry of integrated information. *PLoS Comput Biol* 2009;**5**:e1000462.

Bartels A. Defending the structural concept of representation. *Theoria* 2006;**55**:7–19.

Bayne T, Spence C. Multisensory perception. In: Matthen M (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press, 2015.

Bednar JA, Wilson SP. Cortical maps. *The Neuroscientist* 2015;**22**:604–17.

Block N. Inverted earth. *Philos Perspect* 1990;**4**:52–79.

Block N. Semantics, conceptual role. In: Craig E (ed.), *The Routledge Encylopedia of Philosophy*. London: Routledge, 1998.

Ritchie JB, Carruthers P. The bodily senses. In: Matthen M (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press, 2015.

Brouwer GJ, Heeger DJ. Decoding and reconstructing color from responses in human visual cortex. *J Neurosci* 2009;**29**:13992–4003.

Bruineberg J, Dolega K, Dewhurst J, Baltieri M. The Emperor's New Markov Blankets. *Behav Brain Sci* 2021;1–63. 10.1017/S0140525X21002351.

Chalmers D. *Constructing the World*. Oxford: Oxford University Press, 2012.

Churchland PM. *A Neurocomputational Perspective*. Cambridge: MIT Press, 1989.

Churchland PM. State-space semantics and meaning holism. *Philos Phenomenol Res* 1993;**53**:667–72.

Churchland PM. Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered. *J Philos* 1998;**95**:5–32.

Churchland PM. Neurosemantics: on the mapping of minds and the portrayal of worlds. In: White KE (ed.), *The Emergence of Mind*, Milan: Fondazione Carlo Elba, 2001, 117–47. Reprinted in: P. M. Churchland (2007): Neurophilosophy at Work. Cambridge University Press.

Clark A. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York: Oxford University Press, 2016.

Clark A. *Sensory Qualities*. Oxford: Clarendon Press, 1993.

Clark A. *A Theory of Sentience*. New York: Oxford University Press, 2000.

Collins T, Tillmann B, Barrett FS *et al.* A combined model of sensory and cognitive representations underlying tonal expectations in music: from audio signals to behavior. *Psychol Rev* 2014;**121**:33–65.

Cummins RC. *Meaning and Mental Representation*. Cambridge, MA: MIT Press, 1989.

Cummins RC. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press, 1996.

Decock L. A physicalist reinterpretion of 'phenomenal' spaces. *Phenomenol Cognit Sci* 2006;**5**:197–225.

Diedrichsen J, Kriegeskorte N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput Biol* 2017;**13**:e1005508.

Edelman S. Representation is representation of similarities. *Behav Brain Sci* 1998;**21**:449–67.

Edelman S, Grill-Spector K, Kushnir T *et al.* Toward direct visualization of the internal shape space by fMRI. *Psychobiology* 1998;**26**:309–21.

Fink SB, Kob L, Lyre H. A structural constraint on neural correlates of consciousness. *Philos Mind Sci* 2021;**2**:7.

Fodor JA. *Psychosemantics*. Cambridge, MA: MIT Press, 1987.

Fodor JA, Lepore E. *Holism: A Shopper's Guide*. Cambridge, MA: Blackwell, 1992.

French S. *The Structure of the World: Metaphysics and Representation*. Oxford: Oxford University Press, 2014.

Friston KJ, Fagerholm ED, Zarghami TS *et al.* Parcels and particles: Markov blankets in the brain. *Network Neurosci* 2021;**5**:211–51.

Gärdenfors P. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press, 2000.

Gert J. Quality spaces: mental and physical. *Philos Psychol* 2017a;**30**:525–44.

Gert J. *Primitive Colors: A Case Study in Neo-pragmatist Metaphysics and Philosophy of Perception*. Oxford: Oxford University Press, 2017b.

Gescheider GA. *Psychophysics: The Fundamentals*, 3rd edn. Mahwah, NJ: L. Erlbaum Associates, 1997.

Gładziejewski P. Action guidance is not enough, representations need correspondence too: A plea for a two-factor theory of representation. *New Ideas Psychol* 2016;**40**:13–25.

Gładziejewski P, Miłkowski M. Structural representations: causally relevant and different from detectors. *Biol Philos* 2017;**32**:337–55.

Goodman N. *Languages of Art: An Approach to a Theory of Symbols*, 2nd edn. Indianapolis: Hackett Publishing Company, 1976.

Hardin CL. *Color for Philosophers: Unweaving the Rainbow*. Indianapolis: Hackett, 1988.

Hardin CL. Color quality and structure. In: Hameroff S, Kaszniak A, Chalmers D (eds), *Toward a Science of Consciousness III*, Cambridge, MA: MIT Press, 1999, 65–74.

Haxby JV, Connolly AC, Swaroop Guntupalli J. Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* 2014;**37**:435–56.

Hohwy J. *The Predictive Mind*. New York: Oxford University Press, 2013.

Isaac AMC. Objective similarity and mental representation. *Australas J Philos* 2013;**91**:683–704.

Isaac AMC. Structural realism for secondary qualities. *Erkenntnis* 2014;**79**:481–510.

Janata P. Navigating tonal space. In: Hewlett W B et al. (eds.), *Tonal Theory for the Digital Age (Computing in Musicology 15)*. Stanford, CA: CCARH, 200739–50.

Kim J. Multiple realization and the metaphysics of reduction. *Philos Phenomenol Res* 1992;**52**:1–26.

Kim J. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press, 1998.

Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern* 1982;**43**:59–69.

Kohonen T. *Self-Organizing Maps*, 3rd edn. Berlin: Springer, 2001.

Kohonen T. Self-organized maps of sensory events. *Philos Transact Royal Soc A* 2003;**361**:1177–86.

Koulakov AA, Kolterman BE, Enikolopov AG *et al.* In search of the structure of human olfactory space. *Front Syst Neurosci* 2011;**5**:65.

Kriegeskorte N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci* 2008;**2**:4.

Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition computation, and the brain. *Trends Cognit Sci* 2013;**17**:401–12.

Kuehni RG, Schwarz A. *Color Ordered. A Survey of Color Order Systems from Antiquity to the Present*. New York: Oxford University Press, 2008.

Ladyman J, Lambert J, Wiesner K. What is a complex system? *Eur J Philos Sci* 2013;**3**:33–67.

Lawless HT. *Quantitative Sensory Analysis: Psychophysics, Models and Intelligent Design*. Hoboken, NJ: Wiley-Blackwell, 2013.

Lewis D. Ramseyan humility. In: Braddon-Mitchell D, Nola R (eds), *Conceptual Analysis and Philosophical Naturalism*. Cambridge: MIT Press, 2009.

Loorits K. Structural qualia: a solution to the hard problem of consciousness. *Front Psychol* 2014;**5**:237.

Lyre H. The "multirealization" of multiple realizability. In: Hieke A, Leitgeb H (eds), *Reduction - Abstraction - Analysis*. Frankfurt: Ontos, 2009, 79–94.

Lyre H. Humean perspectives on structural realism. In: Stadler F (ed.), *The Present Situation in the Philosophy of Science*. Dordrecht: Springer, 2010, 381–97.

Lyre H. Verkörperlichung und situative Einbettung. In: Stephan A, Walter S *Hg.: Handbuch Kognitionswissenschaft*. Stuttgart: Metzler, S, 2013, 186–92.

Lyre H. Quantum identity and indistinguishability. In: Friebe C *et al. The Philosophy of Quantum Physics*. Berlin: Springer, 2018a, 73–101.

Lyre H. Newton Goethe and the alleged underdetermination of ray optics. *J Gen Philos Sci* 2018b;**49**:525–32.

Malach R. Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci Conscious* 2021;**7**:1–13.

Marvan T, Polák M. Generality and content-specificity in the study of the neural correlates of perceptual consciousness. *Philos Mind Sci* 2020;**1**:5.

Müller O. Prismatic equivalence – a new case of underdetermination: Goethe vs. Newton on the prism experiments. *Br J Hist Philos* 2016;**24**:322–46.

Newen A, Leon DB, Gallagher S (eds) *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press, 2018.

Newman MHA. Mr. Russell's causal theory of perception. *Mind* 1928;**37**:137–48.

O'Brien G, Opie J. A connectionist theory of phenomenal experience. *Behav Brain Sci* 1999;**22**:127–96.

O'Brien G, Opie J. Connectionist vehicles, structural resemblance, and the phenomenal mind. *Commun Cognit* 2001;**34**:13–38.

O'Brien G, Opie J. Notes toward a structuralist theory of mental representation. In: Clapin H, Staines P, Slezak P (eds), *Representation in Mind: New Approaches to Mental Representation*, Oxford: Elsevier, 2004, 1–20.

O'Callaghan C. Perception and multimodality. In: Margolis E, Samuels R, Stich SP (eds), *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford: Oxford University Press, 2012.

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.

Palmer SE. Fundamental aspects of cognitive representation. In: Rosch E, Lloyd BL (eds), *Cognition and Categorization*. Hillsdale: Erlbaum, 1978.

Palmer SE. Color, consciousness, and the isomorphism constraint. *Behav Brain Sci* 1999;**22**:923–43.

Papineau D. Can we really see a million colours? In: Coates P, Coleman S (eds), *Phenomenal Qualities: Sense, Perception, and Consciousness*. Oxford: Oxford University Press, 2015.

Pearl J. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann Publishers, 1988.

Putnam H. Realism and reason. *Proc Addr Am Philos Assoc* 1976;**50**:483–98. Reprinted in: H. Putnam (1978): Meaning and the Moral Sciences. New York: Routledge.

Quine WV. Two dogmas of empiricism. *Philos Rev* 1951;**60**:20–43.

Raffman D. Similarity spaces. In: Matthen M (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press, 2015.

Ramsey W. *Representation Reconsidered*. Cambridge: Cambridge University Press, 2007.

Ramsey W. Untangling two questions about mental representation. *New Ideas Psychol* 2016;**40**:3–12.

Rang M, Passon O, Grebe-Ellis J. Optische Komplementarität. Experimente zur Symmetrie spektraler Phänomene. *Phys J* 2017;**16**:43–9.

Robbins P, Aydede M (eds) *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press, 2009.

Rosenthal D. How to think about mental qualities. *Philos Issues* 2010;**20**:368–93.

Rosenthal D. Quality spaces and sensory modalities. In: Coates P, Coleman S (eds), *Phenomenal Qualities*. Oxford: Oxford University Press, 2015.

Roskies AL. Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. *Synthese* 2021;**199**:5917–35.

Russell B. *The Analysis of Matter*. London: Allen & Unwin, 1927.

Shagrir O. Structural representations and the brain. *Br J Philos Sci* 2012;**63**:519–45.

Shapiro LA. Multiple realizations. *Journal of Philosophy* 2000;**97**:635–54.

Shea N. Exploitable isomorphism and structural representation. *Proc Aristotelian Soc* 2014;**114**:123–44.

Shea N. *Representation in Cognitive Science*. Oxford: Oxford University Press, 2018.

Shepard RN. Cognitive psychology: review of the book by U. Neisser. *Am J Psychol* 1968;**81**:285–9.

Shepard RN. Geometrical approximations to the structure of musical pitch. *Psychol Rev* 1982;**89**:305–33.

Shepard RN, Chipman S. Second-order isomorphism of internal representations: shapes of states. *Cogn Psychol* 1970;**1**:1–17.

Swoyer C. Structural representation and surrogative reasoning. *Synthese* 1991;**87**:449–508.

Tononi G, Boly M, Massimini M *et al*. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.

Tononi G, Koch C. Consciousness: here, there and everywhere? *Philos Transact Royal Soc London B* 2016;**370**:20140167.

Tsuchiya N, Phillips S, Saigo H. Enriched category as a model of qualia structure based on similarity judgements. *Conscious Cogn* 2022;**101**:103319.

Tsuchiya N, Saigo H. A relational approach to consciousness: categories of level and contents of consciousness. *Neurosci Conscious* 2021;**7**:1–13.

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.

van Fraassen BC. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press, 2008.

von der Malsburg C. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 1973;**14**:85–100.

Willshaw DJ, von der Malsburg C. How patterned neural connections can be set up by self-organization. *Proc Royal Soc London Ser B* 1976;**194**:431–45.

Young BD, Keller A, Rosenthal D. Quality-space theory in olfaction. *Front Psychol* 2014;**5**:1.