PhiMiSci
Philosophy and the Mind Sciences | Vol. 6 | 2025

Structural representation and the Newman problem of the brain and AI

Holger Lyre[1,*]

Special Issue: *Structuralism in the Science of Consciousness*, edited by Sascha Benjamin Fink & Andrew Lee

Abstract

The brain is an operationally closed system. At it's boundary, every incoming signal gets translated into neural activity. The intrinsic nature of the distal causes perturbing the brain at its sensory surfaces cannot be grasped from the inside. Moreover, the brain exploits change detection and relational coding. Hence, the external relations between any distal causes are transformed into internal relations between neural activities. The picture of a thoroughgoing structuralism about the mind emerges: all mental states are relationally individuated and thereby provide structural representations. But structuralism is vulnerable to a rather generic logico-mathematical problem: the Newman problem. The specific "Newman problem of the brain" is the concern that the brain's operational closure is accompanied by a representational closure, where neither the intrinsic nature of the distal causes nor the nature of the external relations can be grasped from the inside. Potentially, this applies to any operationally closed structural representation system, including brains and AI systems. If the Newman problem persists, then such systems "know" nothing, the outside world remains completely hidden. This is nothing less than a skeptical scenario of the most extreme form. However, I will present a solution to this conundrum. It works by "spatio-temporal grounding": spatial and temporal relations are unalteredly transferred from the external world to systems of structural representation.

Keywords: Grounding • Neurophenomenal structuralism • Newman problem • Operational closure • Quality space • Structural representation

[1] University of Magdeburg
[*] Primary contact: lyre@ovgu.de

Lyre, H. (2025). Structural representation and the Newman problem of the brain and AI. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11855

1 Introduction

Suppose you're a neuron. What does the world around you look like? You are connected to other neurons in your vicinity. And those other neurons are connected to still other neurons. And you're all doing the same thing: to spike or not to spike. That's pretty much it. So it is spiking or non-spiking neurons: a pretty impoverished list, indeed, of the building blocks of the brain. And amazingly, out of these building blocks, the brain seems to be able to build representations of the external world. How could that ever be possible?

Suppose, now, you're an electric current in the digital circuits of some AI system. What does the world around you look like? The small electronic circuit in which you get processed, presumably a logic gate, is connected to other gates in the vicinity. And those other gates are connected to still other gates. And you're all doing the same thing: to be on or off. That's pretty much it. So it is electric currents or non-currents: a pretty impoverished list, indeed, of the building blocks of the AI system. And amazingly, out of these building blocks, the AI system is supposed to be able to build representations of the external world. How could that ever be possible?

These scenarios suggest the following observations:

O1 Both brains and AI systems are (or should function as) representational systems.

O2 Brains and AI systems are operationally closed: their building blocks, the basic vehicles of their representational machineries, are of one and the same kind, and the operational boundary of the system is given by the boundary of the building blocks.

O3 There is nothing intrinsic to spikes or digital currents as representational building blocks that they share with their worldly targets.

The first two observations imply a tension. While O1 says that brains and AI systems function as representational systems, O2 portrays such systems as operationally closed. So there's a tension: if a representational system is "closed", how should it ever represent? In order to represent, a representational system must somehow be able to reach beyond the system boundary and to capture whatever it aims to represent. Operational closure must therefore not entail representational closure, but must in principle be compatible with a certain representational 'openness'.

Obviously, the notions of representational system and operational closure need some unpacking. First of all, a representational system R consists of a (set of) physical vehicle(s) that can be attributed a content such that R serves to represent the target T (representandum) of R. Hence, R is individuated by its content and realized by its vehicles.

The notion of closure should be understood here in the mathematical sense of the closure of an algebraic structure. The natural numbers, for instance, are closed under the operations of addition and multiplication (but not subtraction and division), since any addition or multiplication of any pair of natural numbers yields another natural number (therefore, the set of natural numbers forms a semigroup with respect to addition and a semigroup with respect to multiplication). As a matter of principle, the closure of a domain is thus relative to an operation. A domain is operationally closed, if all corresponding operations defined on it remain within the same domain. (We shall see later, in section 5.2, that the notion of operational closure must be properly understood as a functional notion based on a functional understanding of the notion of operation.) Hence, the brain, as observed by O2, is operationally closed under the operation of neural spiking. Here we assume that spiking is the representationally relevant operation (i.e. relevant for the brain to act as a representational system). Moreover, and this is a further indication, the vehicle operations of the brain as a representational system are operationally closed, so that the brain retains its systemic unity and integrity.

However, despite the operational closure (which concerns the vehicle operations of R), the brain should not and cannot be closed as far as the *representation relation* (or "operation") between R and T is concerned. Since T lies outside the brain, the representation relation must reach beyond the vehicle boundary of the system. In this sense, a genuine representational system can never be "representationally closed". In order to represent, R and T must somehow be related. At a minimum, R must possess a causal grounding in T such that R is causally downstream of T. To use Putnam's (1981) vivid example: a configuration created by cosmic accident that looks like Churchill doesn't represent Churchill. In order for R to represent T, there must have been some causal connection between R and T (however weird and tangled and possibly also going back a long way in time).

Causal grounding, however, is at best a minimal condition for representation. This brings us back to the diagnosed tension between the brain's operational closure and the quest for representational openness. Let us augment O1 with the following, mildly realist assumption: to yield a *realistic internal representation* of the external world, the representation must somehow resemble the worldly target. Resemblance or similarity can, in addition, be spelled out in terms of shared properties. Our assumption can thus be reformulated in the following way:

> A *realistic internal representation* must share certain properties with the external target.

Under this assumption, the above tension morphs into a problem. First of all, O3 restricts the required representational openness: representations and targets cannot share intrinsic properties. So if brains and AI systems function as realistic representational systems, the representations that they yield can at best share relational or structural properties with their worldly targets. In other words: brains and AI systems cannot represent by means of intrinsic property similarity, but rather by structural similarity. In still other words: if brains and AI systems function as realistic representational systems, then they must work on the basis of *structural representations*.

But what if the systems' *operational* closure also entails *representational* closure? What if the representational system is *not only unable* to capture the intrinsicality of the targets, but also their relational properties, i.e. the target's structure? In that case, brains and AI systems could just yield arbitrary structural representations, and that would of course violate any realist assumption. This, indeed, is an instance of the so-called *Newman problem* in the special regime of structural representations for brains and AI systems.

The Newman problem is in fact a rather generic logico-mathematical problem. It has its origin in Max Newman's 1928 response to Bertrand Russell's "Analysis of Matter". I present the historical background of the problem as well as its modern incarnation in the debate about structural realism in philosophy of science (mostly physics) in section 2. In our present context, however, we must first ask whether, in the case of brains and AI systems, operational closure and representational closure indeed collapse. For only in that case the Newman problem will occur. In section 3, it will be shown to which extent this is indeed the case. For the sake of exposition I will focus my presentation in this and the following two sections to neural systems only. In section 5, we will extend our perspective and argue that all of the previous findings also apply, mutatis mutandis, to AI systems and extended mind hybrids of neural and artificial systems.

The core part of the paper is section 4, since the Newman problem of the brain leaves us with a skeptical scenario of the most extreme form: brains (and AI systems) "know" nothing, the outside world remains completely hidden. I will argue, however, that a solution to this conundrum can be sought and found. It works by "spatio-temporal grounding": spatial and temporal relations are unalteredly transferred from the external world to neural systems.

My analysis not only provides a solution to the Newman problem, but also leads straight to the concept of structural representations. Structuralism about the mind – with regard to both intentionality and phenomenality – has recently experienced a strong boost. And it turns out to be clearly supported by our considerations in the wake of the Newman problem. This will be discussed at length in section 6. I will conclude, in section 7, with a summary list of the various findings obtained in the course of the paper and open remarks about the grounding of higher-level thinking and scientific theorizing.

2 The Newman problem: historical and systematic background

2.1 Historical background

As mentioned, the Newman problem is a generic logico-mathematical problem that has its historical origin in Bertrand Russell's theory of knowledge and the escalating versions of structuralism that he developed in the 1910s and 1920s. A first version can be found in "The Problems of Philosophy" (1912), where Russell proclaims:

> We can know the properties of the relations required to preserve the correspondence with sense data, but we cannot know the nature of the terms between which the relations hold […] [A]lthough the relations of physical objects have all sorts of knowable properties, […] the physical objects themselves remain unknown in their intrinsic nature. (1912: Chap. 3)

Russell had thus come to the structuralist conclusion that the intrinsic nature of things in the external world is inaccessible to us. This is of course reminiscent of Kant's thesis that we cannot grasp the things-in-themselves. But even if we cannot grasp the intrinsic properties of things, Russell still believed that we can grasp the *nature of the relations* in which things stand. Fifteen years later, however, he also abandoned this idea. In his "The Analysis of Matter" (1927), Russell now claimed that we have knowledge about the world by structural descriptions only:

> [W]herever we infer from perceptions, it is only structure that we can validly infer; and structure is what can be expressed by mathematical logic, which includes mathematics. […] The only legitimate attitude about the physical world seems to be one of complete agnosticism as regards all but its mathematical properties. (1927: 254, 270-271)

This, indeed, is an extreme view, for it means that not only are relations insufficient to determine the intrinsic nature of things, but that the nature of relations themselves remains indeterminate.[1]

One year later, in 1928, the thus appointed young lecturer in mathematics at Cambridge, Max Newman, wrote a paper on "the causal theory of perception", the "theory of our knowledge of the unperceived parts of nature" that Russell had put forward in his 1927 book. Newman made it unmistakably clear that Russell's claim that only structure is known is highly problematic, for on this view

> [t]he world consists of objects, forming an aggregate whose structure with regard to a certain relation R is known, say W; but of the relation R nothing is known (or nothing need be assumed to be known) but its existence; that is, all we can say is, "There is a relation R such that the structure of the external world with reference to R is W". (1928: 144)

However, this has a dramatic consequence, since

> [a]ny collection of things can be organised so as to have the structure W, provided there are the right number of them. Hence the doctrine that only structure is known involves the doctrine that nothing can be known that is not logically deducible from the mere fact of existence, except ("theoretically") the number of constituting objects. (1928: 144)

This is the Newman problem: extreme structuralism is near vacuous, all we can know is just cardinality. We will see in section 4 how Russell responded to this, but end here with a quote that shows what Newman himself envisaged as perhaps the only possible solution to the problem:

> The conclusion that has been reached is that to maintain the view that something besides their existence can be known about the unperceived parts of the world it is necessary to admit direct apprehension of what is meant by the statement that two unperceived events are *causally*

[1] See Demopoulos & Friedman (1985) and van Fraassen (2008, chap. 9) on Russell's path to structuralism.

> *adjoined*, i.e., happen near each other, temporally and spatially, or overlap, or do something of the sort. The central doctrine is then that while of percepts we have a qualitative knowledge, of other events all that can legitimately be inferred is their structure with regard to a certain directly known relation which may be called "causal proximity". (1928: 148)

2.2 Levels of structuralization

Russell upheld different versions of structuralism in the 1910s and 1920s. This is possible since structuralism comes in levels. Our knowledge about the world can be more or less structuralized, and world models can capture their domain of discourse in more or less structural terms. Consider the description or knowledge that a model delivers in terms of a formula in predicate logic. It may contain variables, individual constants, predicates and logical expressions. The predicates in the formula could represent either monadic properties or relations (polyadic properties), and they have variables or individual constants as their arguments. All of this is reminiscent of the machinery of the Ramsey sentence to capture the ontological content of theories and models.

To give an illustrative example, let us consider the central city part of the Sydney train network as our target domain (cf. Figure 1).[2] A full-blown object-oriented ontology of this toy domain consists in a full description including the station names, the train line names and their stops. Without loss of generality, we can just focus on a small section, for example:

> Blue Line with 1. stop at Central & Blue Line with 2. stop at Town Hall & Green Line with 1. stop at Central & Green Line with 2. stop at Museum

We may abbreviate this as:

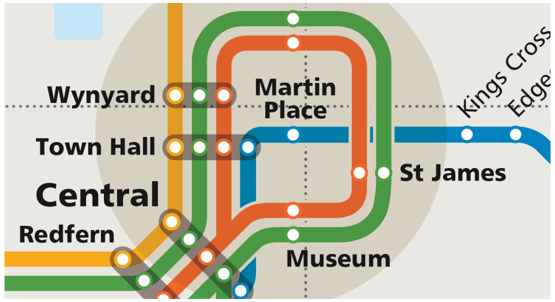

Figure 1: Map of the central part of the Sydney train network

$$B1(c) \;\&\; B2(t) \;\&\; G1(c) \;\&\; G2(m) \qquad (S0)$$

Here we have used predicates B1, B2, G1, G2 and individual constants c, t, m. The predicates can be understood as qualitative and intrinsic properties of the individuals in our domain. The first step of structuralization consists in replacing the individuals by variables and fixing them with existential quantifiers:

$$\exists\, x,y,z\ \{B1(x) \;\&\; B2(y) \;\&\; G1(x) \;\&\; G2(z)\} \qquad (S1)$$

The entities in our domain are no longer maximally individuated, they are individuated by their properties only. They may still possess an intrinsic nature, but they lack individuality beyond properties. Hence, the formula can be satisfied by multiple triples of individuals. This is a general observation: structuralization opens the door to multiple realizability. A second step of structuralization consists in replacing the monadic predicates by relations:

$$\exists\, x,y,z\ \{B(x,y) \;\&\; G(x,z)\} \qquad (S2)$$

Entities in the domain are now not otherwise individuated than by the relations in which they stand. For instance, we know from the above formula that there are two entities that are connected by the Blue Line, one of which is also connected to a third entity by the Green Line. The third and final step

[2] Rudolf Carnap famously used the railway network example in his "Aufbau" (Carnap, 1928). Chalmers (2022) uses the New York City subway system for a similar illustration.

of structuralization consists in replacing even the relations by quantified variables (going over to second-order predicate logic):

$$\exists\, x,y,z \ \exists\, R1,R2\ \{R1(x,y)\ \&\ R2(x,z)\} \qquad \text{(S3)}$$

This is the ultimate step of structuralization. In this step, not only the nature of the relata is dropped, but also the nature of the relations. The formula consists, besides variables, of logical connectives only. This step amounts to the formal method of Ramsification of scientific theories. The Ramsey sentence of a given (finitely axiomatized) theory about the world is obtained by replacing the theoretical predicates by second-order variables and then existentially quantifying over those variables.

2.3 The structural realism debate

If we apply the considerations of section 2.2 to 2.1, we see that Russell in 1912 took the position of an S2-structuralism, while in 1927 he ended in the ultimate S3 version. Only this strong version is prone to the Newman problem, which is, again, a generic problem of S3-structuralism.

In the preceding section, I also spoke rather broadly about knowledge in terms of (structural) world models. But it wasn't specified whether models of perception are considered, or models of lower or higher cognition or even scientific models. Regarding the latter, structuralism has become a most influential and widespread position in the debate about scientific realism under the heading of *structural realism* (cf. French, 2014, Lyre, 2010). But as the focus of this paper is on neural systems and the brain, I will restrict myself to models of perception and cognition.[3]

Ladyman (1998) introduced the useful distinction between epistemic and ontic structuralism to the structural realism debate. Russell's view, for instance, much like Kant's agnosticism about things-in-themselves, is often considered an epistemic version of structuralism (ESR). It tells us what we can and cannot capture about the external world. His view has a mind-to-world direction of fit and is silent about the ontology of the world. Ontic structural realism (OSR) is more ambitious. It is driven by the idea that we should not be silent about ontology but that "structure is all there is". Unfortunately, this much cited slogan is ambiguous. If we think of structures as sets of relata endowed with sets of relations, then the slogan might read as the radical view that there are no relata but only relations. But this is obviously more than questionable. Most proponents of OSR therefore adopt a more moderate interpretation according to which there are relations and relata, maybe even ontologically on a par, but in the sense that the relata are nothing over and beyond the relations in which they stand.

The Newman problem has led to a variety of considerations in the modern structural realism debate, and authors have come up with a variety of views according to which the Newman problem is either restricted to ESR and does not pertain to OSR (Ladyman, 1998), is still unsolved (Ainsworth, 2009) or can be overcome (Melia and Saatsi, 2006; Kleiner, 2024a). The problem is also considered a close cousin of Putnam's (1976) model-theoretic argument.[4] It is the goal of the present paper to add to this debate a brain-focused perspective that is informed by modern neuroscience. I will not be concerned with questions about ontology, but I do want to claim that the proper origin of the Newman problem and the proper way to find a solution is on the level of the "Newman problem of the brain". And this is a fact that, if I am right, has yet been overlooked in the structural realist debate about the Newman problem.

[3] However, as the example of Russell shows, the nature of our models of perception may eventually have an impact on the nature of our scientific models. Russell was interested in a general theory of knowledge and he drew immediate conclusions from his structuralism about perception to scientific knowledge. I will only discuss scientific knowledge in the context of a brief outlook in the concluding section 7.

[4] 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). See also the remarks by Chalmers (2012, chapters 1, 8).

3 The Newman problem of the brain

The Newman problem is a generic problem for structuralism. What does it have to do with the brain? Surprisingly a lot, as quickly becomes apparent when one considers the peculiarities of the brain's system boundary and the mechanisms of the transition from external influence to internal neuronal activity. In this section we will (re)discover the Newman problem of the brain from the ground up, so to speak. For a more compact presentation, I shall use "N" to refer to the domain of neural states and processes (neural activities) and "W" for the domain of worldly states and processes (external stimuli).[5]

3.1 The neural/non-neural boundary

In broadest terms, our world consists of either neural or non-neural systems. They connect at a specific interface that I call the neural/non-neural boundary ("3N", for short). The 3N boundary consists of an inward part W ⟶ N and an outbound part N ⟶ W.

The inward part W ⟶ N is the *boundary of neural transduction*. Here, distal causes affect the neural system at its sensory surfaces. Because of the mechanisms of neural transduction, the inward boundary has a crucial implication: the nature of the incoming signals changes. This means that, whatever the nature of the distal cause perturbing the neural system was gets translated into neural activity and spike trains inside the neural system. Whether electromagnetic, chemical, or acoustic stimuli reached the receptors of the brain's sensory organs, everything gets translated into spikes. From the perspective of the neural system, therefore, the nature of the distal perturbing causes remains hidden.

The outbound part N ⟶ W is the *boundary of neuromotor transformation*. Here, efferent motor neuron activity transforms into bodily actions of effector organs, mostly muscles and glands. At this outbound part we 'leave' the N-domain, so to speak. But the same rationale as for the inward part applies: what happens outside of N remains hidden from the inside. Hence, whenever motor neuron firing gets translated into muscle contraction or, more generally, into behavior, the neural system reaches beyond its 3N boundary and thereby produces effects in a hidden W-domain.

Taken together, the two parts of neural transduction and neuromotor transformation combine to the full 3N boundary. Nevertheless, our focus in the following will be on the inward boundary. We saw that from the way this boundary works everything gets translated into one single "currency": neural activities. As stated in the introduction, a domain is operationally closed, if all corresponding operations defined on it remain within the same domain. In this sense, the neural domain of the brain is *operationally closed*. Another way to spell this out is to say that the 3N boundary covers or shields the internal neural states of N from the external non-neural states of W. From the point of view of statistical learning or machine learning, the 3N boundary can also be viewed as a Markov blanket (see Lyre, 2022, Footnote 3).

3.2 Change detection and relational coding

Because of the 3N boundary, the nature of latent external signals cannot be transferred from W into N. The 3N boundary leads to operational closure of the neural system, and there is concern that it even leads to representational closure. To test this, we need to understand at least the basic principles of how the transition from W to N works at the sensory surfaces.

Whenever the neural system gets perturbed, N undergoes some change at its sensory periphery. And that change can potentially be detected. For instance, a photon of a certain energy hits a photo receptor 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.

Of crucial importance is the term *change*. Neural systems are, at the boundary of neural transduction, sensitive to changes in the environmental

[5] This section draws heavily on Lyre (2022), where I initially presented the Newman problem of the brain in the context of neurophenomenal structuralism (see section 6.2).

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.

In ontological terms, change detection means the following: regardless of whether the causal stimuli of the world can be assigned intrinsic or relational properties, only relational properties will be transferred over the 3N boundary. It is fairly easy to see what these relational properties on the level of the neural system amount to: 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 or temporal coding, sparse or dense 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 relational properties of the coding elements.

To sum up: the brain's 3N system boundary and the mechanisms of change detection and relational coding give rise to an operational and maybe even representational closure of the brain. The latter, however, is the threat of the *Newman problem of the brain*. The point of the problem is, again, the following: not only do relations not suffice to pick out the intrinsic nature of the objects in the domain, but also the nature of the relations themselves remains undetermined. As far as we have introduced and discussed the structure of N, the perceiving and thinking subject is threatened by the Newman problem. From the perspective of the neural system, nothing can be 'known' about the external W-causes perturbing N. At least that is how it seems

4 Solving the Newman problem of the brain

Is there any solution to the Newman problem? We encountered the problem on two different routes: as a general logico-mathematical problem of structuralism and as a problem that arises specifically for the brain due to its given boundary, detection and coding conditions. In this section I present a solution to the Newman problem. More precisely: by bringing the two routes together, it becomes clear that the real origin of the Newman problem of structural realism lies in the Newman problem of the brain and that a solution to *this* problem can be sought and found.

4.1 The Newman-Russell exchange

Remarkably, a solution to the Newman problem was indeed already found, or at least hinted at, by Russell himself. But the story is a little strange. First of all, Newman's objection had an immediate impact on Russell. Already in April 1928 he sent a letter to Newman with the following response:

> 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.

This is a clear fallback from S3 to S2, Russell's earlier 1912 version of structuralism. Or at least, it hints at a position that sits somewhat between S2 and S3. Why? Because still not *all* relations are fixed or grounded, but only spatio-temporal relations. What makes the story strange is that Russell makes a clear admission to Newman in his letter, but never returns to the whole topic in his later publications. One reason may be (but I'm only speculating here) that he in retrospect thought that he *had* already sufficiently hinted at his "spatio-temporal continuity with the world of percepts” in his “Analysis of Matter”. For example, in a passage, where he first repeats that “our knowledge of physics is mathematical … because no non-mathematical properties of the physical world can be inferred from perception”, but then adds:

> There is, however, one exception to this limitation, at least apparently. The exception I mean is *time*. We always assume that the time between percepts is the same as the time in the physical world. (1927: 253)

Although the immediately following paragraphs qualify this statement a little, here is then a later passage:

> A piece of matter is a logical structure composed of events; the causal laws of the events concerned, and the abstract logical properties of their spatio-temporal relations, are more or less known, but their intrinsic character is not known.

A charitable interpretation may read this as the claim that, in Russell’s terminology, we are ‘directly acquainted’ with spatio-temporal relations. But as Russell exegesis is not my task here, I will rather leave the final word to Newman who rightly points out:

> The conclusion that has been reached is that to maintain the view that something besides their existence can be known about the unperceived parts of the world it is necessary to admit direct apprehension of what is meant by the statement that two unperceived events are causally adjoined, i.e., happen near each other, temporally and spatially, or overlap, or do something of the sort. The central doctrine is then that while of percepts we have a qualitative knowledge, of other events all that can legitimately be inferred is their structure with regard to a certain directly known relation which may be called “causal proximity”. (1928: 148)

4.2 The spatio-temporal grounding of N in W via perception

This, I think, is indeed the road to the solution of the Newman problem of the brain. A solution that is in tune with our modern understanding of the brain: ultimately, our knowledge of the world is spatio-temporally grounded. And taking into account the regularity of space-time sequences, we can perhaps even say that we have access to the causal structure of the world, as Newman points out. But I will restrict myself here to the spatio-temporal structure (even only to topological spatio-temporal structure, as we will see below).

In the light of section 3, our consideration can be stated as follows: Perception is based on change detection, and neural systems work on a purely relational basis. While it seems that neither the intrinsic nor the relational W-properties can be conveyed over the 3N boundary, in fact a sufficient part of spatio-temporal W-relations carry over to N-relations. In more detail: (i) the spatial separation of two stimuli on a sensory surface gets transferred as two spatially separated neural activations, and (ii) the temporal sequence of the stimuli gets transferred as a temporal sequence of neural spikes. We can therefore directly refer to such spatio-temporal W-relations, we are equipped with a *spatio-temporal grounding of N in W via perception*:

> *Whatever the nature of the external stimuli, the spatio-temporal proportions of the stimuli can (but need not) be conveyed over the 3N boundary.*

Let me illustrate this with a few concrete examples (cf. Lyre, 2022). Consider two successive tactile stimuli at two different spots on your arm. Clearly, the nature of the stimuli, the mechanical force, remains ‘unknown’ to the neural system, as it gets transduced into neural activity. But the spatio-temporal proportion of the stimuli can 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. Now take, 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 hidden to the neural system, but the spatial as well as the temporal proportions of the stimuli are directly sensible. Finally, 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.

4.3 Sensory intersections

Humans are, of course, equipped with more than just one sensory system and, moreover, none of our sensory systems works in isolation. Different sensory organs overlap or intersect in important ways. These facts support the spatio-temporal grounding in many ways.

Let us start with a small thought experiment. Imagine a cognitive system or subject X with only one sensory channel. X has reason to assume that there are external, worldly causes of a hidden nature, whose changes can be detected by its sensory channel. It may therefore infer that there is more "out there" than mere space-time structure. Now suppose that X has more than just one sensory channel. Under the *independence assumption* that different sensory channels are sensitive to stimuli of different nature, if X has *n* sensory channels, then X may assume that there are *n* types of external causes in W. In other words, while X cannot capture the particular nature of the external causes, neither their intrinsic nature nor the nature of their relations in W, it may make guesses about the cardinality of the hidden natures.

But such guesses are highly fragile, since the above independence assumption is most probably false. There are no strictly independent sensory channels. In any case, the independence assumption cannot be checked from within the system. To start with, what is a sensory channel? Neither receptor cell types nor (certainly not!) sensory organs fulfill the requirements. The reason is that all of them partially overlap in some of their functions.

Consider the eye as a sensory organ for vision. The eye itself is clearly no independent sensory channel since there are four different receptor cells in the retina. They can be grouped into cones and rods. But this amounts neither to two nor to four independent channels, since we can find partially overlapping response curves for all cells. So there is generally a lot of cross-talk between biological sensory systems, which is one reason why the longstanding debate about the number of senses cannot be settled in a unique fashion. And it is also the reason why the brain can at best provide a rough estimate about the cardinality of the external hidden natures in W.

While this is a skeptical limitation, the various sensory overlaps and intersections also have a positive and supportive function for the spatio-temporal grounding. Some sensory receptors can for instance be used in multiple ways for different detection purposes. Retinal receptors are not only receptive to visual but also to (strong) mechanical stimulation, which may lead to cloud-like visual impressions. Some temperature sensors in the mouth respond to "hot" spices the same way they do to hot temperatures. And very bright light may be painful and, in audition, loudness or high pitch may be painful as well.

Moreover, 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. I can feel a sugar cube on my tongue and it tastes sweet. And I can hear a deep bass tone and feel it in my stomach. These sensory intersections serve to support each other in calibrating our various senses and relating them onto each other in a systematic and orchestrated fashion. This is important for matters of grounding. Since none of our sensory systems works in isolation, the neural system inside the 3N boundary operates as an integrated whole. And the spatio-temporal grounding of N in W thereby infiltrates and pervades the entire neural system. This leads to a grounding of the entire system.

However, we must add one final qualifying remark. All the arguments in favor of a solution to the Newman problem of the brain and its spatio-temporal grounding support the assumption that spatial and temporal neighborhood relationships can cross the 3N boundary. This gives us access to the topology of space and time, but not to the metric structure. The spatio-temporal grounding of the brain is at best a topological grounding.[6]

5 The Newman problem of AI and more

The Newman problem of the brain arises since the brain works as a representational system that is confined by the 3N boundary. The 3N boundary has two effects: it leads to operational closure, but at the same time suggests the additional possibility or danger of representational closure. Briefly stated, the Newman problem of the brain lies in the question of whether the representational closure is indeed complete.

We can think of a system as a structured domain or organization that consists of components or vehicles performing certain system-specific operations. A system is operationally closed, if all the internal operations defined on the vehicles of the system's domain remain within that same domain. In this way, the system retains its integrity and organizational unity. In the case of the brain, the candidate vehicles are neural activities: the electrical and chemical processes that occur within neurons, notably spikes or spike trains. Due to the 3N boundary, such activities are constrained by change detection and relational coding. And this, as we have seen, means there is nothing intrinsic to neural activities that they share with worldly goings-on, but that only the relational properties of the neural activities can be exploited for representational purposes. Call the totality of the relational properties of neural activities the *relational structure of the neural vehicles*, or, *neural structure*, for short. And call the totality of the relational properties of the external world the *world structure*. The Newman problem of the brain then lies in the question of whether and to what extent internal neural structure corresponds to external world structure.

5.1 Extending the Newman problem

Clearly, the Newman problem of the brain is not restricted to the *human* brain. In fact, we could speak more precisely and at the same time more generally of the Newman problem of the nervous system, at least the central nervous system (CNS, considered as a representational system). And as such it applies to all organisms with a CNS in the animal kingdom. So this is a first step of generalizing the problem, but even more interesting is the question whether we can generalize the Newman problem beyond neural tissue.

The short answer is yes. And the best way to see this is to give the Newman problem its most general formulation in terms of the Newman problem of structural representation. This formulation can almost directly be derived from the above:

> *Newman problem of structural representation*: Given a structural representational system S and its operational closure, will the representational closure of S be complete?

This shows that the Newman problem is actually a problem of structural representation, and we have more to say about the notion of structural representation in the following section. For now we characterize structural representations by the idea that the dependency relation between a representation and its target is the relation of *structural similarity*. 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 are

[6] Topological grounding is in line with at least three different and general considerations that can be linked here (two of which were suggested by two anonymous reviewers):

Within the N-domain, the brain, the literal spatial distance between neurons is, of course, largely irrelevant. What matters is the connection distance in the network. This is consistent with topological grounding: the spatial distance of external stimuli in W is first and foremost an indicator of their sheer numerical distinctness. It transforms into distinct but typically somehow connected neural signals. Thus, at best, spatial neighborhood is conveyed.

There are abundant distortions and illusions of time and space perception (e.g. the tactile cutaneous rabbit illusion). Again, in such cases, at best the topological but not the metric structure of the stimuli is conveyed.

Prima facie, our considerations are in harmony with spacetime conventionalism, i.e. the view that metrical geometry is not a fixed, real property of physical spacetime. But of course this goes far beyond the scope of the current paper.

maps, pictures and sculptures. Structural similarity can be either first-order or second-order (cf. Shepard and Chipman, 1970; O'Brien and Opie, 2004). In the case of first-order structural similarity the corresponding relations have the same nature, in the case of second-order similarity they do not.[7] Hence, a structural representational system S is representationally closed, if it exploits all and only second-order structural similarity. But to overcome the Newman problem, S must go beyond mere second-order similarity and be able to grasp the nature of *at least some* of the external relations in terms of first-order structural similarity.

To avoid being misunderstood: second-order structural representations are of great advantage. It's a real virtue that we can use weather maps without getting wet or being blown away, since what is represented in those maps differs in nature from the medium of representation. But in this case the grounding of the weather map is done by us, external interpreters and users, of the map. Such simple interpretivism is not on offer when it comes to the Newman problem of the brain. Here, the solution lies in the spatio-temporal grounding of neural structure in world structure, hence, in the first-order structural similarity between spatial and temporal relations on the two sides of the 3N boundary. This qualifies the Newman problem as a special problem for representational systems that are "autonomous" in the sense that their "representational power" arises, as it were, out of themselves. The usual understanding is that this is the case for the human mind and likewise for animal minds and, potentially, also for artificial minds. So the question arises: Is there a Newman problem of AI?

As already indicated in the introduction, there is a clear analogy between neural systems and modern machine learning or AI systems. First, such systems show operational closure. They are based on a defined set of vehicles: electric currents in electronic circuits. This is a result of their system boundary (the "electric/non-electric boundary", as it were). Moreover, the boundary works such that the systems rely on change detection and relational coding. Electronic sensors get triggered by external physical changes only, and inside the system only the relational changes of currents may be exploited for the purpose of representation. But much like neural spikes, electric currents "live" in space and time. Spatially separated external causes trigger spatially separated (parts of) electronic sensor surfaces as much as temporally separated causes trigger temporally separated parts. And they lead to spatially and temporally separated electric currents or signals inside an AI system.

The Newman problem is therefore not limited to brains or neural systems, analogous problems arise for all confined signal domains that serve the purpose of autonomous representation as for instance electric signals in AI systems. In all such cases, a grounding of such systems and thereby a solution to the Newman problem can only be guaranteed by their spatio-temporal grounding.

5.2 Extended Mind: Extending the 3N boundary

A natural next question to ask is: what happens when biological and artificial systems get combined? Neural implants would be a first example. Consider the already existing cases of retinal and cochlear implants. Their medical benefit lies in compensating for damage in the receptor cells by having artificial electronic receptor units take over their function and being connected to downstream cells in the sensory organ. What do we have to conclude about the overall system boundary in such cases? What about the validity of the 3N boundary? Shouldn't we say, as most people would say, that the neural implant becomes part of the "neural" system and that the system boundary includes the implant? And likewise for future neural implants that might substitute internal parts of the brain? The answer is a clear yes and it shows that the term "3N boundary" (and analogous terms like "electric/non-electric boundary") has indeed a misleading element in it. It does not mean that our previous considerations about the 3N boundary

[7] Cf. Lyre (2022, 5) for examples of second-order structural similarity: In a bar chart, rectangular bars or columns are used with heights proportional to the data that they represent. In a weather map, the spacing of isobars corresponds to pressure gradients in the atmosphere. There are also mixed forms of first- and second-order similarities. A true-to-scale road map represents spatial distance relations in nature by corresponding spatial distance relations on the map, but the map is made of a different material than the objects it represents.

were wrong. But if we consider more general cases than pure biological brains and neural systems, our definition of the system boundary has to be refined. We can see this from the general formulation of the Newman problem of structural representation in the previous section. The term 3N boundary doesn't occur here. What counts is that the representational system shows *operational closure*. We have seen that a system is operationally closed, if the corresponding operations defined on the vehicles of the system's domain remain within that domain. We now add that operational closure is, ultimately, a functionally defined notion, since the notion of operation is a functional notion. The retinal implant takes over the *function* of the damaged receptor cells in the retina. It gets seamlessly integrated into the overall mechanism of the retina. Most notably, the electronic components of the retinal implant, which take over the receptor function, operate according to the same functional principles as the neuronal receptor cells: they use change detection and relational coding.

In other words: From a functionalist perspective, a more general understanding of structural representational systems (and the Newman problem) arises. In fact, structural representation and functionalism go hand in hand. The functionalist picture of operational closure is this: operations of the vehicles of mental representations are of the same functionally individuated kind. The operations of artificial neural implants and biological neurons are of the same kind, since they fulfill the same function. The vehicles and their operations must be individuated in a functional and largely medium-independent way. This is in tune with what we have already seen: that the intrinsic nature of the representing vehicles doesn't play a role and cannot be exploited by any system for its representational purposes.

On the face of it, the term "3N boundary" looks medium-dependent. This is misleading. The 3N boundary is in fact a functionally defined boundary of natural biological neural systems and brains. In the case of the neural implant the system becomes slightly extended. Since the extended system (brain plus implant) provides an operationally closed whole, a total system of functionally integrated mechanisms, the new system boundary is given by the boundary of this bigger whole.

This opens the door to the celebrated *extended mind thesis* (Clark, 2008). And it brings our considerations about the Newman problem of structural representation in harmony with this thesis. But of course, the thesis goes much further than cases of neural implants. Suppose we replace the neural implant by some (not too futuristic) implanted sender-receiver device that is connected over, say, Bluetooth to some external computing system. Then most of the function of the artificial vehicle is now provided outside of skin and skull. Clark & Chalmers (1998) argue that for reasons of parity the external device should now be considered a genuine part of the extended cognitive system. This is their well-known parity principle which states that if external vehicles are functionally equivalent to vehicles inside the head, and if those latter vehicles are regarded as cognitive, then the external vehicles should be regarded as cognitive as well. This reasoning is clearly functionalist in spirit, as is the extended mind thesis in general. And this is the reason why cases of mind extension, here understood as operationally closed hybrids of neural and artificial systems including externally coupled artificial systems (with a high enough and functionally equivalent bandwidth) are threatened by the Newman problem in much the same way as cases of pure neural or pure AI systems. They show operational closure with a functionally defined system boundary, and that boundary creates a Newman problem. Hence, such extended systems have to draw on change detection and relational coding and will be representational systems that exploit structural representations. And sure enough, their grounding has to be a spatio-temporal grounding.

6 Structural representations

The solution to the Newman problem, at the minimum a partial solution, is provided by spatio-temporal grounding. Brains and AI systems are grounded in the world since they can grasp the nature of spatial and temporal relations. In a sense, this is almost a triviality: since representational systems are themselves physical systems in space and time, it must be the case that whatever vehicles are used inside for the purpose of coding and representation, these vehicles have spatio-temporal proportions. But there is another insight that comes equipped with our solution that is by no means trivial. There seems to be really no other way for brains and AI systems to represent than by means of structural representations. This has already become apparent in the previous section. In the case of the brain, only neural *structure* – relational properties of the neural vehicles – can be exploited for representational purposes.

Recall that the Newman problem only occurs for structural representation. But there were two routes by which we encountered the problem. The first was the 'high road' from Newman's original abstract consideration to modern structural realism debates and the solution that structural realism offers to the problem of scientific realism and scientific representation. The second route, the one that we took in section 3, was to encounter the Newman problem as it arises specifically for the brain due to its given boundary, detection and coding conditions. This yields the following line of argument: the specific boundary, detection and coding conditions of the brain result in a Newman problem. But the problem only occurs for structural representations. Hence, given its specific conditions, the brain exclusively instantiates structural representations.

This provides a nice argument for structuralism about the brain. And given the generalizations laid out in the previous section, the point generalizes beyond brains to AI systems and extended hybrids. They all exclusively instantiate structural representations. With this in mind and for the sake of brevity, I will again limit myself to considerations of the brain. And in the very next subsection I will present even another argument in favor of structural representations. The argument will be based on the long-overlooked fact that, according to modern approaches, the brain frequently works in a *generative* rather than merely passive and representational mode. But for a model to be a suitable *generative model*, it must have a structural similarity to the world. In short: generative models must be structural representations.

6.1 Structural versus indicator notions of representation

Before we delve into structural representations, let us first consider their counterpart: *indicator representations* (also receptor or detector representations). Indicator representations do not represent by means of (any sort of) similarity, but due to their indicator function. And therefore, their representational vehicles must represent intrinsically (rather than relationally or structurally). In more general terms, the distinction between indicator and structure representations can be captured by the difference in the *representational dependency relation* d(R,T) that the two accounts postulate between a representation R and its target T. The indicator account of representation postulates that d(R,T) is "R functions as an indicator of T", while structural representationalism postulates that d(R,T) is "R is structurally similar to T". The indicator account is highly fashionable both within neuroscience and connectionism, the doctrine behind artificial neural networks and, hence, modern AI. It goes back to the pioneering work of Hubel & Wiesel (1959) on edge detecting neurons in V1 and culminates in the concept of so-called grandmother neurons. If the brain (or AI systems) indeed works on the basis of indicator representations or can be grounded in their workings, then the Newman problem would not arise, as it only arises for structural representations. But we have already made it sufficiently clear that such a problem exists. It is therefore necessary to critically discuss and assess the indicator notion of representation.[8]

Consider edge detecting neurons. Isn't it obvious that they provide low-level representations? Indeed, many connectionists take the indicator notion of representation to be part and parcel of their view. But unnecessarily so. Connectionist systems, unlike classic symbolic systems, are built on sub-symbols and the vehicles of such sub-symbols are regarded as feature detecting neurons. In other words: the neurons in the primary stages of our cognitive hierarchy – in our terminology: neurons that are part of or close

[8] The distinction between structural and indicator representations has been criticized in the literature by Morgan (2014) and Nirshberg & Shapiro (2020), but defended by Cummins & Poirier (2004), Gładziejewski & Miłkowski (2017), O'Brien (2015).

by the inward part of the 3N boundary – just *are* feature representations. But given our findings about the brain's boundary, detection and coding conditions, this can hardly be the case. Specific neurons, for example in the primary visual cortex V1, can correlate stably with specific features, for example with edges in the visual field. But the neural activity in and of itself is not a representation. There is nothing intrinsic to neural activity that can be taken as an intrinsic ingredient of a representation. Several authors have therefore (for different reasons) pointed out that we must carefully distinguish between indicators and representations (cf. Gładziejewski and Miłkowski, 2017, O'Brien 2016). As Cummins & Poirier (2004, 23) put it: "Indicator signals are arbitrary in a way that representations are not. […] anything can be made to indicate anything else". Moreover:

> Indicators are source dependent in a way that representations are not. The cells studied by Hubel and Wiesel all generate the same signal when they detect a target. You cannot tell, by looking at the signal itself (the spike train), what has been detected. […] Indicator signals "say" their targets are present, but "say" nothing about them; representations provide structural information about their targets, but do not indicate their presence. Indicator signals say, "My target is here," while representations say, "My target, wherever it is, is structured like so." (Cummins and Poirier, 2004, 24-25)

Let us turn to grandmother neurons, the pinnacle of the indicator notion. They are said to be highly selective cells that respond to only a few objects in the world, or, in the extreme case, to only one complex object. Grandmother neurons are high-level instances of indicator representations. They are high-level regarding their high-level semantic content. A number of arguments against grandmother neurons have been put forward in the debate. First and foremost, any extreme local coding scheme leads to a combinatorial explosion: it's simply impossible to use a single neuron for every object in the world. After all: how many are they? How to individuate them? Does my grandmother in sunlight count as a different entity than grandma in the evening? Even though the number of neurons in the human cortex is insanely high, the number of (possible) objects in the world still far exceeds this number. Even more compelling is the fact that there exist no reports of "grandmother-specific" deficits or capacity losses! Neural damage can lead to the loss of broad, but not very narrow recognition or memory capacities.

Now consider, for the sake of illustration, a simple artificial multi-layer feed-forward network that has been trained to distinguish images of human faces (including grandma). On the input layer, the network units correlate with – and indicate – simple image features: edges, colors, intensities etc. On the next higher layer the features consist of odd combinations of the features from the level below. And this recipe works for the rest of the multi-layer hierarchy. On the top layer, finally, we get output units that indicate the different faces in the training set (if all goes well). Doesn't provide the top layer grandmother neurons then? And isn't this also demonstrated by neuroscientific evidence in the form of „Jennifer Aniston cells" (Quian Quiroga, 2012)? A first point to notice is that the top layer really is a mere readout or internal indicator, the causal work of the network has already been done on all of the levels below (Thomas and French, 2016). If the network is combined with a behavioral output mechanism – for instance a hug when Grandma appears – then this mechanism could be fed directly from the penultimate level. A second point is that Jennifer Aniston cells provide no sufficient evidence for highly selective grandmother neurons, but somewhat more broadly selective "concept cells" only (Quian Quiroga, 2012).

An intriguing, third point is the following. What would it mean to *generate* a grandmother representation? An intuitive picture is to run the hierarchy of the above network backwards in order to generate the initial input data. This, in simplified terms, is the concept of an encoder-decoder-network. The idea to use neural networks as generative models rather than discriminative models, for instance as a decoder rather than an encoder, lies at the heart of the recent *generative turn* in machine learning (cf. Lyre, 2024). And from this generative perspective the idea of highly selective local neural representations becomes even more obscure. It is absurd to assume that the activity of just one neuron could generate the mental image of a whole person or complex object! Suppose I am recalling my long-dead grandmother. How else could this rich generative model that my brain so marvelously produces come about but through the orchestrated activity of a whole population of neurons? Who would seriously believe that such a mental image could be produced by the activity of a single neuron? Instead, two things should have become clear. The brain exploits population coding, and the generative models carried by these neuronal populations must *resemble* their targets. Since, however, only neural *structure* – relational properties of the neural vehicles – can be exploited, such resemblance must be structural. Generative models thus provide a further striking argument in favor of structural representations.[9]

6.2 Structuralism about the mind

In recent years, the structural representation (S-Rep) account has become an increasingly popular account of mental representation, and the analyses and results put forward in this paper provide clear and additional support for this account. Pioneering work was done by Cummins (1989, 1996), Palmer (1978), Swoyer (1991) and, most notably, Gerard O'Brien & Jonathan Opie (2004).[10] Central to the concept of structural representation are the notions of structural similarity and structure-preserving mappings between representation R and target T.

The criterion of structural similarity makes S-Rep a (moderate, as only structural) version of the classical (and crude) similarity account of representation. The crude account, however, is subject to well-known objections: The similarity relation is symmetrical, the representation relation is not (my passport photo represents me, but I do not represent my passport photo); no matter how similar twins may be, one never represents the other; under weak conditions of similarity, everything becomes similar; conversely, similarity seems to depend on the observer or context. To rebut these kinds of objections, the new, more sophisticated S-Rep account postulates three qualifying specifications or conditions for structural representations.

A first, straightforward fix is to replace the unnecessarily strict but often used criterion of an isomorphism (a bijective homomorphism) for the structure-preserving mapping between R and T with homomorphism in general, which then allows for non-symmetric mapping relations as a general requirement for representations (Bartels, 2006). Even more important are the two further conditions of causal grounding and exploitability.

In a remarkable paper, Alistair Isaac (2013) has introduced the notion of natural representations that builds on objective similarity and where one structure represents another only if it is causally downstream from it. Here, the legacy of the causal theory of reference and causal covariance theory of representation becomes apparent: for R to be a representation of T, there must exist some causal path from T to R, however winding. A randomly created image of Churchill doesn't represent Churchill (to use Putnam's 1981 well-known example). Random homomorphy is not a sufficient criterion for representation. The spatio-temporal grounding to which structural representational systems such as the brain can be traced back meets precisely the causal grounding condition. Spatio-temporal grounding is ipso facto causal grounding.

The third crucial condition is that structural representations must be usable or exploitable. A number of authors have argued for this (Isaac, 2013, Shea, 2014, 2018, Gładziejewski and Miłkowski, 2017), but already Cummins (1996) has observed that the distinction between content and use opens the door for explaining misrepresentation. More precisely, as Isaac (2013) notes, misrepresentation reduces to misuse, since natural representations cannot misrepresent.

The new, eleborated S-Rep account is naturalistic in a down-to-earth fashion, as structural representations are *vehicle representations*, i.e. physical entities by their very nature. The causal work behind their potential use lies fully in their vehicle structure. The account has also a pragmatic or

[9] This result fits nicely to recent work that emphasizes an understanding of generative models in predictive processing accounts as structural representations: Gładziejewski (2015), Kiefer & Hohwy (2018, 2019), Wiese (2017) and Williams (2017).

[10] Further S-Rep supporting contributions comprise Bartels (2006), Isaac (2013), Gładziejewski & Miłkowski (2017), Lee & Calder (2023), Miłkowski (2023), Plebe & De La Cruz (2018), Piccinini (2022), Ramsey (2007), Shagrir (2012), Shea (2014, 2018), Williams & Colling (2018) as well as the references from the previous footnote. For critical voices see Artiga (2023) and Facchin (2024).

instrumentalist flair to the extent that successful use and exploitation are normative notions that depend at least partially on pragmatic ascriptions.

So far we have related structuralism to intentionality, but structuralism has recently also been applied to phenomenality. Kleiner (2024b) even speaks of a "structural turn in consciousness science". The key idea here is that the specific qualitative content of phenomenal experiences consists in the structural facts that are systematically encoded in the totality of all possible sensory discriminations. This makes phenomenal content a relational affair and accounts for structural representations encoded in quality spaces. Pioneering work was done by Austen Clark (1993, 2000) and David Rosenthal (2010, 2015).[11]

Following this work, Fink et al. (2021) and Lyre (2022)[12] have proposed *neurophenomenal structuralism* (NPS) as an agenda for a structuralist neuroscience of consciousness. NPS rests on the two assumptions that (i) any phenomenal experience is fully individuated by its place in a quality space structure, and that (ii) quality space structure is mirrored in neural structure. We can think of NPS as a weak version of structural representation with self-organized neural maps as vehicles that mirror quality space structures and also represent feature structure of the world (Lyre, 2022; more on this below). A special feature is that NPS leads to *phenomenal holism*, since each internal quality space relation is determined by the whole quality space structure. NPS thus declares that dichromats, for instance, experience all colors differently, since their quality space structure differs from the quality space structure of trichromats.[13]

It is instructive to distinguish three types of structural representations concerning three different types of structures involved (Lyre, 2022):

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 regarding the spatial structure of both representation and target. In the case of the brain, a striking example is provided by place cells in the hippocampus, which operate as a map based on location-related activity (cf. Shea, 2018, Chap. 5.2). 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. There are also mixed cases of type-1 and type-2. But in all such cases the structural representations mirror concrete spatio-temporal affairs of the world and will thereby rely on spatio-temporal grounding. From a neurocomputational point of view, such cases will be implemented by population coding as part of a state-space-semantics in the sense of Churchland (1989, [2001): relationships in content are encoded in the distance relationships of neuronal population states in the activation space. Hence, the metric structure of the state space of such neural structural representations has a semantic interpretation.

The third type of neural representations is given by *neural maps*. Such type-3 structural representations are the typical neural vehicles for quality space structures. Neural maps are ubiquitous in the brain, the best known class are cortical maps (cf. Bednar and Wilson, 2015). Empirical evidence

[11] Further contributions comprise O'Brien & Opie (1999, 2001), Gärdenfors (2000), Gert (2017), Davies (2021), Decock (2006), Edelman (1998), Fleming & Shea (2024), Isaac (2014), Lee (2021), Loorits (2014), Malach (2021), Palmer (1978, 1999), Shepard (1968) and Tsuchiya et al. (2021, 2022).

[12] See also Kob (2023) and Fink (2024).

[13] Note that NPS differs from Chalmers' "phenomenal structuralism" (Chalmers, 2012, chap. 8.7). The latter is a version of structuralism in the spirit of Carnap's Aufbau that includes only structurally defined phenomenal properties as base truths. As Chalmers notes this even "leaves open a panpsychist version of phenomenal structuralism, on which one in effect specifies the properties of microphysical entities by specifying the total experiences of those entities in terms of their phenomenal structure" (Chalmers, 2012, 416). This is almost the opposite to NPS where phenomenality via its quality space structure is grounded, if not reduced, by the mirroring assumption in neural and, hence, physical structure.

for all types of structural representations comes from multivariate pattern analysis, most notably RSA: representational similarity analysis (cf. Kriegeskorte, 2008, Kriegeskorte and Kievit, 2013).

6.3 Grounding phenomenal content

In the light of the Newman problem, type-3 representations are of special interest. They are parasitic on the spatio-temporal grounding of type-1 and type-2 representations, since they do not directly mirror concrete spatio-temporal affairs in the external world. Their grounding is only guaranteed, if at all, by sensory intersections (as pointed out in section 4.3). Rather than mirroring concrete spatio-temporal world structure, type-3 representations mirror *hidden world structure*. The structure is hidden or latent as it pertains the external domains of relations that the brain assumes to exist beyond the domains of space and time according to its various sensory channels. Mutatis mutandis, this includes the kind of properties that traditional epistemology has called "secondary qualities". While primary qualities such as size, shape, motion, solidity and number can all be, more or less directly, grounded in spatio-temporal properties, secondary qualities give rise to phenomenality. According to NPS, they must be construed as structural representations about the sensory content conveyed by the corresponding sensory channels. Moreover, such structural representations are encoded in quality space structures with a mirroring in neuronal structures.

Two important observations can be made about "secondary qualities". First, they come equipped with both *content* and *character*. To say that a mental state has phenomenal character is to say that there is something it is like to be in that state. Phenomenal character can be understood as *experienced content* (Lyre, 2022). It is then a plausible option for NPS to consider phenomenal character as a composite of specific content and a general mechanism that makes that content what-it's-like. NPS has nothing to say about this general mechanism, but it has a lot to say about phenomenal content: the content is fully individuated by its place in a quality space structure.[14]

The second, even more important observation is that phenomenal contents are *not independent from each other*. Our color judgments, for instance, stand in similarity relations: orange is between yellow and red and is the opposite of blue. Similarity ratings among perceptual stimuli can then be conceived as delineating the structure of quality spaces. This is the gist of any structural approach about phenomenality. Phenomenal contents depend on the relations in which they stand. This is what specifies and individuates content. And structuralism about phenomenality provides the perfect fit to the empirical fact of phenomenal similarity ratings.

It is typically overlooked – surprisingly enough! – that this likewise constitutes a strong argument against the traditional qualia conception. For this view can in and of itself *not* account for the empirical fact of phenomenal similarity ratings. The reason is that qualia enthusiasts consider qualia as intrinsic. An ascription of intrinsicality to something is entirely about that something, and it is irrespective of the way anything else is. The specific phenomenal character of a specific quale is therefore independent and obtains irrespective of any other qualia. But this means that, for the qualia enthusiast, the systematic phenomenal similarities must come out as sheer coincidences!

To elaborate on this, note that qualia cannot come in degrees. Any quale is qualitatively different from any other quale. There is no quale of color in general, but specific color qualia. And this holds true on any level of specificity or coarse graining. Even if we could only distinguish four basic colors in human color phenomenality (say yellow, red, blue and green), then because of their intrinsicality nothing determines how to order them.[15] It is, however, an empirical fact that humans order colors by their

[14] In the parlance of Marvan & Polák (2020), NPS is about the neural correlate of content (NCc) rather than the general neural correlate of consciousness (gNCC).

[15] An anonymous reviewer made the following objection: Rest mass is a fairly good example of an intrinsic property. But obviously mass comes in similarity relations. A mass of 10 kg is more similar to a mass of 11 kg than it is to a mass of 100 kg. The corresponding ratings are no coincidences. While this is true, the objection

systematic similarities. This is a crucial part of color phenomenality itself. The qualia enthusiast must accept this as a brute fact, he cannot account for similarities, while the structuralist naturally embraces them.[16]

So we said that structuralist accounts of phenomenality such as NPS understand phenomenal content on the basis of type-3 structural representations encoded in quality spaces. These considerations were mostly based on the first assumption of NPS (that any phenomenal experience is fully individuated by its place in a quality space structure). Now what about the claim that type-3 representations mirror hidden world structure? Here, the second NPS assumption comes into play, according to which quality space structure is mirrored in neural structure.

Consider again the example of color and color spaces. According to our findings in section 4.3, we cannot grasp the nature of (what we call) “color”, not even the nature of purportedly existing relational color properties. But surely we have color experiences. Therefore, colors are most likely a posit of the human mind. Our mind “represents” such posits as structural representations encoded in color spaces. But why call them representations at all? Certainly, they do not directly correspond to spatial or temporal relations (they are neither type-1 nor type-2 representations). But they are no arbitrary fictions either. Secondary qualities like colors are indirectly related to the external world. Here is a helpful quote from Rosenthal (2010, 378) that we can use as a springboard for our considerations:

> The spatial properties perceptible by different sensory modalities are of course the same; the physical shapes, sizes, and locations we perceive by sight are the same as those we perceive by touch. But the corresponding mental qualities are not. Vision determines spatial perceptible properties as boundaries of color, whereas taction determines them as boundaries pertaining to perceptible pressure and texture. The mental qualities that pertain to spatial properties are special to each of the sense modalities. Cross-modal calibration of the spatial properties discerned by each modality must be learned.

The spatio-temporal grounding consists in learning the cross-modal calibration of perceptual boundaries. Such boundaries are the spatial and temporal boundaries of otherwise unknowable, hidden causes that perturb the sensory surfaces of our various sensory organs. The brain has no grasp on the nature of such hidden causes, nor on the nature of the relations between these causes. But it may well grasp their spatial and temporal boundaries.

And, as Rosenthal rightly notices, such cross-modal calibration must be learned. This means that the neuronal structures that mirror quality space structures are only developed through *adaptive* processes and *learning* processes. In other words, the neural structures adapt to world structures in such a way that these neural structures can be exploited for all sorts of relevant purposes and functions (survival, reproduction, homeostasis, perception, navigation, memory, thinking etc.). Although the details of these adaptive and learning developments are still largely unknown, the following consideration is sufficient for our purposes: *in order to form a neural map, it is economical for the cognitive system if stimuli that occur more frequently in temporal or spatial proximity are also represented internally in proximity*. This is the gist of self-organized neural maps (SOMs) and their development (Kohonen, 2001). The neural vehicles of quality spaces are such SOMs. This explains why, for instance, the arrangement of hues in the human color space corresponds at least roughly to the arrangement of colors in the rainbow.[17]

nevertheless rests on a misunderstanding. Of course, mass comes in grades, and such grades immediately allow for similarity ratings. But again: a color quale doesn’t come in grades. Each individual color is an intrinsic quale. In other words, a quale is a determinate, not a determinable. Two particular colors, say green-35 and blue-29, correspond to two particular and different color qualia. This should be clear: green and blue aren’t just grades of a single (and weird) green-blue-quale. Or even more extreme: all of the (millions of) colors aren’t just grades of one (master?) color quale. Rather, the qualia enthusiast has to assume that there are millions of individual and intrinsic color qualia with no natural connections to one other.

[16] Of course, this doesn’t rule out the *metaphysical possibility* of the strong qualia position. The metaphysical possibility remains, as usual, as the last (albeit incredibly implausible) refuge of the strong position. See my analogous remarks on the issue of qualia inversion in Lyre (2022).

[17] See Lyre, 2022 for further remarks on SOMs and dimension reduction, which is why, for instance, the linear sequence of frequencies in the prism spectrum and

To expand on this explanation: physics tells us that light is to be understood as electromagnetic waves in the visible part of the electromagnetic spectrum, and the ordering of (what we see as) colors in the rainbow is due to the refraction of sun light waves in raindrops. Of course, the term "electromagnetic wave" is just another, high-level scientific term hinting at unknown hidden causes. But whatever electromagnetic phenomena are (see the concluding section 7), they *also* have spatial and temporal properties and boundaries. Indeed, most of the physical behavior of electromagnetic waves stems from the spatio-temporal properties of such waves: their wave-"length" and their frequency (oscillations per "time" interval). The neural map for color in V4 tracks these spatio-temporal properties, since in natural worldly environments, electromagnetic waves with similar frequencies (or wavelengths) occur much more frequently in spatial proximity than abrupt frequency jumps. In this way, type 3 representations are indirectly related to the outside world.

7 Conclusion: Newmanian skepticism and scientific realism

The Newman problem of the brain proves to be a rich and complex issue, and we went a long way from its historical origin over its diagnosis as an intricate problem for the brain and other operationally closed representational systems to its proposed solution and consequences in terms of the concept of structural representation. Here is a summary of the various findings along the way:

(1) Structuralism comes in levels S1-S3, only the ultimate S3 version is prone to the Newman problem.

(2) Neural systems and brains are operationally closed, the 3N boundary covers or shields the internal neural states from the external non-neural states of the world.

(3) The brain uses the mechanisms of change detection and relational coding.

(4) (2) and (3) motivate the Newman problem of the brain. It is the worry that not only the intrinsic nature of the distal causes cannot be grasped from the inside but also the nature of the external relations.

(5) The Newman problem of structural representation reads: given a structural representational system S and its operational closure, will the representational closure of S be complete?

(6) Spatio-temporal grounding provides a solution to the Newman problem: whatever the nature of the external stimuli, the spatio-temporal proportions of the stimuli can be conveyed over the 3N boundary.

(7) No sensory system works in isolation, the brain operates as an integrated whole, and the spatio-temporal grounding thereby infiltrates and pervades the entire neural system.

(8) Because of (3) only neural structure can be exploited for representational purposes. The Newman problem of the brain is the Newman problem of structural representation.

(9) A structural representational system is representationally closed, if it exploits all and only second-order structural similarity.

(10) The Newman problem arises for all confined signal domains in the service of autonomous representation such as AI systems and extended hybrids.

(11) A new argument for structural representations in the brain: (2) and (3) result in a Newman problem, but the problem only occurs for structural representations. Hence, the brain exclusively instantiates structural representations.

(12) A further argument for structural representations in the brain: grandmother cells cannot generate rich high-level representations, but only generative models that structurally resemble their targets.

rainbow is mapped onto a color wheel-like structure, where the purple hues (not contained in the spectrum!) are *added* to connect blue and red.

(13) Structural representations are vehicle representations, the causal work behind their potential use lies fully in their vehicle structure.

(14) We can distinguish three types of structural representations involving either spatial, temporal or feature structure.

(15) The traditional qualia conception considers any specific quale as intrinsic and independent from any other qualia, and can therefore not account for the empirical fact and phenomenality of similarity ratings.

(16) The content of phenomenal experiences is fully individuated by its place in a quality space structure. This is the gist of any structural approach about phenomenality including neurophenomenal structuralism (NPS).

(17) Neural maps provide type-3 structural representations and are the typical neural vehicles for quality space structures.

(18) Type-3 neural structural representations indirectly mirror hidden world structures, since their spatio-temporal boundaries provide an indirect grounding.

This is an impressive list. What comes next? Indeed, the Newman topic opens the door to big questions in epistemology and metaphysics. Let me conclude with some final and, evidently, open thoughts.

We can think of the Newman problem as a form of skepticism, a sort of *Newmanian skepticism* about the hidden structures of the world. While classic skepticism typically proclaims that we have no knowledge about the external world, since we miss the intrinsic nature of things, Newmannian skepticism claims that we cannot even know structure, since we miss the nature of the relations as well. Fortunately, we have found a solution: our knowledge, based on neural structural representations, is grounded in the spatio-temporal structure of the world. The solution is, I claim, at least good enough to protect us from strong circularity. The accusation of circularity could be made since we have looked at the Newman problem from a perspective that already presupposes some form of grounded thinking. But since we found a positive solution, we created a kind of internal consistency and so the circularity objection can at least be weakened

But still: isn't our solution perhaps only a partial solution? What about the hidden world structures beyond spatio-temporal structure? Are there any at all? Maybe we cannot know even that? We already stumbled across these problems in our discussions in sections 4.3 and 6.3. There we noted that the brain can only conjecturally assume causes whose nature goes beyond the nature of spatio-temporal structures, but that such hidden causes have spatio-temporal boundaries that provide an indirect grounding for our type-3 neural structural representations (which thus indirectly mirror hidden world structures).

As a defense against skepticism, this already sounds weak. But things get even more intricate if we compare the Newman problem of the brain with the Newman problem of structural realism. While the former addresses directly observable, perceivable world structure and our neural representations thereof, the latter addresses non-observable world structure as stated by science (including fundamental physics). The reason is that science, fundamental science in particular, is typically about the unobservable. Accordingly, scientific realism is realism about the unobservable, and can well be distinguished from common sense realism about the perceivable world around us. In other words: while the Newman problem of the brain raises skepticism about common sense realism, the Newman problem of structural realism raises skepticism about scientific realism – even about the moderate version of structural realism (notably OSR). But of course, the two issues connect. That was already Russell's concern. In section [2.3], I even made the claim that the origin of the Newman problem of structural realism and the proper and overlocked way to find a solution is on the level of the Newman problem of the brain. Here is the rough picture behind it: our mind latches onto the world through the window of perception. From simple states of perception, our thinking rises hierarchically to high levels of cognition, including the ability to do mathematics or philosophy and to develop scientific theories about the world.

Note, however, that on higher levels of thinking we seem not to refer to the same levels of reality as on the lower perceptual levels. On the contrary, high-level thinking allows us to extract high-level patterns and regularities of the world, which then provide the basis for a fundamental scientific picture of the world that underlies our everyday experience. The realist's hope, perhaps his last refuge, could be that this will lead to a fruitful self-sustaining circle, where science establishes the naturalistic foundations of our knowledge, including the neural understanding of our cognitive apparatus, on the basis of which it is itself only possible in the form of high-level thinking. But all sorts of open questions remain: How is non-observable structure grounded in observable structure? What is the connection between the contents of perception and the contents of high-level thinking (including theorizing in physics)? Does the spatio-temporal grounding via perception suffice to defeat Newmannian skepticism?

These are all big questions. And I am far from even daring to answer them, since obviously these questions far exceed the scope of a single paper. But I hope I could make it clear that the Newman problem of the brain is a relevant starting point for these questions, that the spatio-temporal grounding plays a crucial role in solving them, and that a structuralist perspective on the issue of representation is essential. If I had succeeded in doing this, much would already have been achieved.

References

Ainsworth, P. M. (2009). Newman's objection. *The British Journal for the Philosophy of Science*, *60*(1), 135–171. https://doi.org/10.1093/bjps/axn051

Artiga, M. (2023). Understanding structural representations. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/728714

Bartels, A. (2006). Defending the structural concept of representation. *Theoria*, *55*, 7–19. https://doi.org/10.1387/theoria.550

Bednar, J. A., & Wilson, S. P. (2015). Cortical maps. *The Neuroscientist*, *22*(6), 604–617. https://doi.org/10.1177/1073858415597645

Carnap, R. (1928). *Der logische Aufbau der Welt*. Weltkreis-Verlag.

Chalmers, D. (2012). *Constructing the world*. Oxford University Press.

Chalmers, D. (2022). *Reality+: Virtual worlds and the problems of reality*. W. W. Norton & Co.

Churchland, P. M. (1989). *A neurocomputational perspective*. MIT Press.

Churchland, P. M. (2001). Neurosemantics: On the mapping of minds and the portrayal of worlds. In K. E. White (Ed.), *The emergence of mind* (pp. 117–147). Fondazione Carlo Elba.

Clark, A. (1993). *Sensory qualities*. Clarendon Press.

Clark, A. (2000). *A theory of sentience*. Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*, 7–19. https://doi.org/10.1093/analys/58.1.7

Cummins, R. C. (1989). *Meaning and mental representation*. MIT Press.

Cummins, R. C. (1996). *Representations, targets, and attitudes*. MIT Press.

Cummins, R. C., & Poirier, P. (2004). Representation and indication. In H. Clapin (Ed.), *Representation in mind: New approaches to mental representation* (pp. 21–40). Elsevier.

Davies, W. (2021). Colour relations in form. *Philosophy and Phenomenological Research*, *102*(3), 574–594. https://doi.org/10.1111/phpr.12679

Decock, L. (2006). A physicalist reinterpretion of 'phenomenal' spaces. *Phenomenology and the Cognitive Sciences*, *5*, 197–225. https://doi.org/10.1007/s11097-005-9012-4

Demopoulos, W., & Friedman, M. (1985). Bertrand Russell's the analysis of matter: Its historical context and contemporary interest. *Philosophy of Science*, *52*(4), 621–639. https://doi.org/10.1086/289281

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467. https://doi.org/10.1017/S0140525X98001251

Facchin, M. (2024). Maps, simulations, spaces and dynamics: On distinguishing types of structural representations. *Erkenntnis*, *90*, 2743–2764. https://doi.org/10.1007/s10670-024-00831-6

Fink, S. B. (2024). How-tests for consciousness and direct neurophenomenal structuralism. *Frontiers in Psychology*, *15*, 1352272. https://doi.org/10.3389/fpsyg.2024.1352272

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*, 7. https://doi.org/10.33735/phimisci.2021.II.70

Fleming, S. M., & Shea, N. (2024). Quality space computations for consciousness. *Trends in Cognitive Sciences*, *28*(10), 896–906. https://doi.org/10.1016/j.tics.2024.06.001

French, S. (2014). *The structure of the world: Metaphysics and representation*. Oxford University Press.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press.

Gert, J. (2017). Quality spaces: Mental and physical. *Philosophical Psychology*, *30*(5), 525–544. https://doi.org/10.1080/09515089.2017.1325451

Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, *193*, 559–582. https://doi.org/10.1007/s11229-015-0783-4

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy*, *32*, 337–355. https://doi.org/10.1007/s10539-017-9577-0

Hubel, D., & Wiesel, T. M. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Isaac, A. M. C. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy*, *91*, 683–704. https://doi.org/10.1080/00048402.2012.743954

Isaac, A. M. C. (2014). Structural realism for secondary qualities. *Erkenntnis*, *79*(3), 481–510. https://doi.org/10.1007/s10670-013-9501-9

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387–2415. https://doi.org/10.1007/s11229-017-1385-0

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, & P. Calvo (Eds.), *The Routledge companion to the philosophy of psychology* (2nd). Routledge.

Kleiner, J. (2024a). The Newman problem of consciousness science [Preprint]. https://philarchive.org/rec/KLETNP-4

Kleiner, J. (2024b). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*, 103653. https://doi.org/10.1016/j.concog.2023.103653

Kob, L. (2023). Exploring the role of structuralist methodology in the neuroscience of consciousness: A defense and analysis. *Neuroscience of Consciousness*, *2023*(1), 1411. https://doi.org/10.1093/nc/niad011

Kohonen, T. (2001). *Self-organizing maps* (3rd). Springer.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Ladyman, J. (1998). What is structural realism? *Studies in History and Philosophy of Science*, *29*(3), 409–424. https://doi.org/10.1016/S0039-3681(98)80129-5

Lee, A. Y. (2021). Modeling mental qualities. *The Philosophical Review*, *130*(2), 263–298. https://doi.org/10.1215/00318108-2021-0007

Lee, J., & Calder, D. (2023). The many problems with S-representation (and how to solve them). *Philosophy and the Mind Sciences*, *4*. https://doi.org/10.33735/phimisci.2023.9758

Loorits, K. (2014). Structural qualia: A solution to the hard problem of consciousness. *Frontiers in Psychology*, *5*, 237. https://doi.org/10.3389/fpsyg.2014.00237

Lyre, H. (2010). Humean perspectives on structural realism. In F. Stadler (Ed.), *The present situation in the philosophy of science* (pp. 381–397). Springer.

Lyre, H. (2022). Neurophenomenal structuralism. a philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1). https://doi.org/10.1093/nc/niab035

Lyre, H. (2024). "Understanding" AI: Semantic grounding in large language models [Preprint]. https://arxiv.org/abs/2402.10992

Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of Consciousness*, *7*(2), 1–13. https://doi.org/10.1093/nc/niab023

Marvan, T., & Polák, M. (2020). Generality and content-specificity in the study of the neural correlates of perceptual consciousness. *Philosophy and the Mind Sciences*, *1*(II), 5. https://doi.org/10.33735/phimisci.2020.II.61

Melia, J., & Saatsi, J. (2006). Ramseyfication and theoretical content. *British Journal for the Philosophy of Science*, *57*, 561–585. https://doi.org/10.1093/bjps/axl020

Miłkowski, M. (2023). Correspondence theory of semantic information. *British Journal for the Philosophy of Science*, *74*(2), 485–510. https://doi.org/10.1086/714804

Morgan, A. (2014). Representations gone mental. *Synthese*, *191*, 213–244. https://doi.org/10.1007/s11229-013-0328-7

Newman, M. H. A. (1928). Mr. Russell's causal theory of perception. *Mind*, *37*, 137–148. https://doi.org/10.1093/mind/XXXVII.146.137

Nirshberg, G., & Shapiro, L. (2020). Structural and indicator representations: A difference in degree, not in kind. *Synthese*, *198*, 7647–7664. https://doi.org/10.1007/s11229-019-02406-6

O'Brien, G. (2015). How does mind matter? Solving the content causation problem. In T. Metzinger & J. M. Windt (Eds.), *Open mind*.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, 127–196. https://doi.org/10.1017/S0140525X99001792

O'Brien, G., & Opie, J. (2001). Connectionist vehicles, structural resemblance, and the phenomenal mind. *Communication and Cognition*, *34*(1-2), 13–38. https://doi.org/10.1080/09515080120083487

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representation* (pp. 1–20). Elsevier.

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. L. Lloyd (Eds.), *Cognition and categorization*. Erlbaum.

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*, *16*, 846979. https://doi.org/10.3389/fnbot.2022.846979

Plebe, A., & De La Cruz, V. M. (2018). Neural representations beyond "plus X". *Minds and Machines*, *28*, 93–117. https://doi.org/10.1007/s11023-017-9446-1

Putnam, H. (1976). Realism and reason. *Proceedings and Addresses of the American Philosophical Association*, *50*, 483–498.

Putnam, H. (1981). Brains in a vat. In *Reason, truth, and history* (pp. 1–21). Cambridge University Press.

Quian Quiroga, R. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597. https://doi.org/10.1038/nrn3251

Ramsey, W. (2007). *Representation reconsidered*. Cambridge University Press.

Rosenthal, D. (2010). How to think about mental qualities [Philosophy of Mind]. *Philosophical Issues*, *20*, 368–393. https://doi.org/10.1111/j.1533-6077.2010.00187.x

Rosenthal, D. (2015). Quality spaces and sensory modalities. In P. Coates & S. Coleman (Eds.), *Phenomenal qualities*. Oxford University Press.

Russell, B. (1912). *The problems of philosophy*. Williams & Norgate.

Russell, B. (1927). *The analysis of matter*. Allen & Unwin.

Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science*, *63*, 519–545. https://doi.org/10.1093/bjps/axr038

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society*, *114*, 123–144. https://doi.org/10.1111/j.1467-9264.2014.00370.x

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Shepard, R. N. (1968). Cognitive psychology: Review of the book by U. Neisser. *American Journal of Psychology*, *81*, 285–289. https://doi.org/10.2307/1420870

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17. https://doi.org/10.1016/0010-0285(70)90006-3

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*, 449–508. https://doi.org/10.1007/BF00485050

Thomas, E., & French, R. (2016). Grandmother cells: Much ado about nothing. *Language, Cognition and Neuroscience*, *32*(3), 342–349. https://doi.org/10.1080/23273798.2016.1242747

Tsuchiya, N., Phillips, S., & Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition*, *101*, 103319. https://doi.org/10.1016/j.concog.2021.103319

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *7*(2), 1–13. https://doi.org/10.1093/nc/niab023

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, *16*(4), 715–736. https://doi.org/10.1007/s11097-016-9473-x

Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*, *28*(1), 141–172.

Williams, D., & Colling, L. (2018). From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese*, *195*, 1941–1967. https://doi.org/10.1007/s11229-016-1196-2