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

Methodological structuralism and the two-factor approach

Implications for consciousness science and AI

Lukas Kob[1,*]

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

[1] Department of Philosophy, OvGU Magdeburg
* Primary contact: lukas.kob@posteo.de

Kob, L. (2025). Methodological structuralism and the two-factor approach: Implications for consciousness science and AI. *Philosophy and the Mind Sciences, 6.* https://doi.org/10.33735/phimisci.2025.11760

Abstract

Methodological structuralism is a research program that seeks to identify neural correlates of consciousness (NCCs) by mapping phenomenal similarity relationships onto the similarity relations between neural population activity. This paper presents a discussion of the potential benefits of methodological structuralism for the neurosciences of consciousness, namely as a specific theory of neural content encoding. In order to achieve this, I supplement it with a metatheoretical framework concerning the relationship between content and consciousness: the two-factor interaction view. Although structuralism provides a comprehensive description of the neural encoding of content, it is inadequate for fully explaining the conscious experience of contents. The majority of current theories of consciousness posit the existence of an additional mechanism that underlies the conscious experience of content. Consequently, if structuralism is indeed correct, progress in consciousness science can be achieved by investigating the interactions between neural mechanisms responsible for consciousness and structures in neural population code activity accounting for the structure of contents. This also has significant implications for consciousness in AI. I discuss these implications, as well as potential empirical avenues for investigating the interaction between content structures and consciousness with cutting-edge neuroscientific methodologies.

Keywords: AI • Content-specific NCC • NCC • Neurophenomenal structuralism • Quality space • Structuralism

1 Introduction

In Fink et al. (2021) my colleagues and I propose neurophenomenal structuralism as the conjunction of two key claims. First, we argue that qualia—the properties of subjective experience (Nagel, 1974)—are fundamentally relational in nature, meaning structural properties individuate qualia. Second, we claim that there is a structural similarity between the relational configurations of qualia and underlying neural activity structures (See also P. M. Churchland, 1986; Clark, 2000). Accordingly, we formulate a structural constraint on neural correlates of consciousness (NCCs) that requires each potential NCC to structurally mirror the structural properties of the experiential content encodes.

In his article, Kleiner (2024) advocates a structural turn in consciousness science but also raises several criticisms, primarily targeting the structuralist assumptions within Fink et al.'s framework. He differentiates between *structuralist* and *structural* approaches. *Structuralist* approaches, according to Kleiner, take seriously the first assumption of neurophenomenal structuralism—the structural nature and individuation of qualia (see also Loorits, 2014; Lyre, 2022; Tsuchiya & Saigo, 2021). This view assumes certain mathematical properties of quality spaces, their neural encodings, and metaphysical claims regarding the "structural nature" of qualia.

In contrast, Kleiner's *structural* approach is less restrictive; it asserts only that conscious qualities can be described mathematically, without committing to a particular mathematical framework or to the idea that qualia are intrinsically relational. As Kleiner (2024, p. 6) points out: "[M]ore often than not, mathematical structure does not individuate individuals." Thus, structural agendas can be applied without claiming that structure defines individuals. While the structuralist view offers a specific stance on the nature of qualia and their neural encoding, the structural approach merely posits that qualia can be captured by some form of mathematical formalism.

This paper aims to reinforce the *structuralist* program by showing how methodological structuralism, when interpreted as a specific theory of content encoding, can significantly advance the science of consciousness. The key for advancing consciousness science lies in distinguishing between neural correlates of content and general neural correlates of consciousness (Fink & Kob, 2024; Marvan & Polák, 2020). Structuralism then plays a critical role by providing a theory on the neural correlates of content that allows for novel research programs targeting the interaction between content structure and consciousness.

I begin with a review of the structuralist encoding model presented in The structuralist theory of phenomenal content. This is followed by a section entitled Remarks on the notion of "content". In the fourth section, Limits and prospects of structuralism in the study of consciousness, I argue that structuralism must be related to theories of consciousness. The fifth section, The one-factor/two-factor framework, then presents a meta-theoretical framework for discussing the relationship between consciousness and content, structurally understood. In the sixth section I discuss Implications for AI consciousness. The final sections, Empirical considerations and Limitations of the structuralist research program, explore ways to investigate the consciousness-content relationship with structuralist methodologies.

2 The structuralist theory of phenomenal content

Contemporary neuroscience has made significant progress in mapping the structures of subjective experience onto the population codes of neural systems. A notable example is the mapping of the color wheel onto state spaces in the visual cortex (Brouwer & Heeger, 2009, 2013). When subjects are presented with color stimuli in fMRI, the similarity structure of the resulting neural activation patterns reflects the similarity structure of the color space. For instance, (Brouwer & Heeger, 2009) demonstrated that a two-dimensional principal component space (Jolliffe & Cadima, 2016) derived from V4 activation reveals a circular color structure corresponding to eight equally saturated colors from the lightness plane 75 in CIELAB space (CIE, 1986). This implies for example that neural patterns related to red are more similar to those associated with orange, than to those elicited

by green. Furthermore, model-based analyses confirmed these findings, leading the authors to suggest that the neural encoding of color structure involves a structural mapping from the perceptual color space to the neural activation space of V4 (Brouwer & Heeger, 2009, 2013).

Studies employing different methodologies (e.g. Bohon et al., 2016; Cichy et al., 2019) have produced similar findings, reinforcing the idea that sensory content is encoded through structural mappings to neural population code space (Jazayeri & Afraz, 2017). While the precise mechanistic basis of these mappings remains incompletely understood, it is widely suggested that these pattern similarity structures provide insights into how visual information is neurally encoded (for a critical review see Roskies, 2021). Specifically, they indicate that similar processing channels are engaged for similar content (Cohen et al., 2015, 2017; Kieval, 2022; Kob, 2023; Kriegeskorte & Diedrichsen, 2019; Kriegeskorte & Kievit, 2013). When combined with other methods, such as pattern component modeling (Diedrichsen et al., 2018) and decoding (Haynes, 2015)—which focus on different aspects of neural information (Kriegeskorte & Diedrichsen, 2019)—evidence for the encoding of specific content in distinct brain areas can be further substantiated. Investigating structural mappings between content and neural activation space is consistent with the connectionist notion that neural systems exhibit content properties because their state space structure mirrors the structure of stimuli (P. S. Churchland & Sejnowski, 1999; P. M. Churchland, 1986; Clark, 2000; Laakso & Cottrell, 2000; Malach, 2021).

Philosophically, a structural view of neural content differs from other accounts, such as the indicator view of content, which posits that correlated activity between a neural system and stimulus types is sufficient to represent that stimulus type (Shea, 2018). According to the indicator view, content emerges independently of any reflection of the relational properties of stimulus space; neural firing is considered intrinsically informative due to its reliable covariance with external events (Dretske, 1997). In contrast, the structural encoding theory demands that the *structure* of the encoded stimulus space be reflected in some aspect of the representational system (Bartels, 2005; R. Cummins, 1995; Gładziejewski & Miłkowski, 2017; Isaac, 2013; A. Y. Lee et al., 2023; J. Lee & Calder, 2023; Morrison, 2020; O'Brien & Opie, 2004; Ramsey, 2007; Shagrir, 2012; Shea, 2023; Swoyer, 1991).

However, recent work suggests that indicators can also be organized into arrays that reflect the structure of the stimulus (Artiga, 2023; Facchin, 2021; Morgan, 2014; Nirshberg & Shapiro, 2021). According to this view, the distinction between structural and indicator representations is not essential. This questions what a structuralist theory of content encoding is supposed to add to existing discussions.

I will answer this challenge by pointing out an essential difference between structural and indicator accounts: structural representations are holistic, whereas indicators are not. I will then elaborate on the three core claims of structuralism which in fact functions as an umbrella term that covers a range of different positions. While it follows that phenomenal structuralism and representational structuralism can be separated in principle, combining both perspectives allows structuralism to be construed as an overarching framework for the mind sciences.

Holism in structural representations arises from the claim that their content is defined only up to isomorphic (or homomorphic) descriptions (Bartels, 2006). As a result, each node within a structural representation can only be identified by its relational properties within the whole structure, implying a holistic and relational constitution of mental content. While indicator arrays may exhibit structural organization (Shea, 2023), they would be processed by downstream systems only as individual indicator relations, rather than in relation to the whole structure to which they belong. In contrast, each component of a structural representation derives its meaning from its relations to the entire representational structure. Thus, structural representations must be *exploited* holistically by downstream systems (Paßler & Doerig, 2024). Structural representations require holistic processing, while indicators do not. Given that exploitability is a critical feature of representations, variations in the way content is exploited imply different types of representations (R. C. Cummins & Poirier, 2004; Gładziejewski & Miłkowski, 2017; Shea, 2018).

A broader structuralist framework would therefore include the requirement of holistic exploitability in addition to structural mirroring and

structural individuation outlined in Fink et al. (2021). Therefore structuralism in the realm of the mental might be broadly defined as the conjunction of three hypotheses:

i. *Structural individuation*: a mental state can be unambiguously individuated by its relational properties.

ii. *Structural mirroring*: structural properties of the mental can be mapped onto structural properties of a lower explanatory level (e.g. neural activation structure; more on that below).

iii. *Structural exploitability*: mental properties are exploited by subsequent systems qua their structural relations to other mental properties.

It is worth noting, however, that neuroscience has not yet established that the similarity structures within population codes are processed holistically by subsequent systems in downstream processing. Nevertheless, some widely used methods implicitly assume a holistic approach when comparing neural activations across individuals. In this framework, connectionist theories propose that differences between individuals can be captured as second-order dissimilarities within overall similarity space structures (Charest et al., 2014; Haxby et al., 2020; Kriegeskorte, 2008; Laakso & Cottrell, 2000). This suggests an underlying assumption that the overarching representational structure is essential for the interpretation of each individual activation pattern within each individual. Such methods could be described as structurally oriented in a methodological sense (Kob, 2023). However, many of today's techniques are logically compatible with non-structuralist views of representational content. As is often the case, the philosophical discussion remains undetermined by (neuro)science.

Let us now turn from representational structuralism to phenomenal structuralism (A. Y. Lee, 2024; see also A. Y. Lee, 2021), before discussing their interrelationship. The philosophical framework of neurophenomenal structuralism at its core, as originally proposed by Fink et al. (2021), argues for a holistic approach to *phenomenality*: phenomenal holism. A key proposition of this structuralist view is that each phenomenal property can be individuated by its relations to other phenomenal properties. For example, the phenomenal experience of "red" is not determined in isolation by an "intrinsic", "atomistic" or "monadic" property, but in relation to all other color experiences. Although color perceptions may appear as monadic or atomistic properties, they are relational properties in reality (Clark, 2000; Rosenthal, 2015, 2025). Accordingly, structuralism in the philosophy of perception claims that phenomena such as color constancy can be grounded in stable relational patterns across color experiences, rather than intrinsic properties of colors (Davies, 2021, 2022; Morrison, 2020). Neurophenomenal structuralism construes phenomenality as holistic and relational (Fink et al., 2021).[1]

Holism has been a subject of extensive debate in discussions of semantic meaning (Block, 2016; P. M. Churchland, 1993; P. M. Churchland & Churchland, 1998; Fodor & Lepore, 2004). While there are several criticisms of holism, compelling counterarguments have been presented and many of arguments do not apply in the same way to phenomenal properties as they do to semantic properties (see also Fink et al., 2021; Kob, 2023; Lyre, 2022). A central criticism of holism is that it introduces instabilities into a system: if a relation changes, the properties of all related individuals would be affected, since each is defined by the overall relational pattern. However, this global connectivity also allows information to be quickly distributed throughout the system. In addition, the magnitude of this effect can vary depending on the specific relational structure of the system. This links structuralism with graph-theoretic approaches that describe specific connectivity patterns that give rise to different emergent properties. Especially small-world networks, that balance information flow and stability, have been discussed in cog-

[1] It should be noted, however, that David Rosenthal (Rosenthal, in press, p. 10) explicitly rejects quality space holism, arguing that just noticeable differences (JNDs) as "individuating relations are [...] highly local." I lack space to discuss the intricacies related to the implications of using either a JND or a similarity-based approach to constructing psychophysical spaces (Wagenaar, 1975). If Rosenthal is right, then highly local versions of structuralism are possible. However, arguments against holism can be countered. In fact, in my view, holism is more of a virtue than a vice (cf. Fink et al., 2021; Kob, 2023; Lyre, 2022).

nitive neuroscience (Watts & Strogatz, 1998). The claim here is not that quality spaces exhibit small-worldness, but considering a graph-theoretic framework shows that holistic systems can in principle achieve stability through different relational configurations.

Holism also suggests interindividual differences in phenomenal qualities, as variations in neural wiring and personal learning history plausibly shape each individual's quality space. Unlike differences in semantic meaning, these interindividual differences in phenomenality are less problematic, since perceptual similarities need only be sufficient to coordinate human behavior (Fink et al., 2021; Kob, 2023; Lyre, 2022). Nevertheless, such differences are often large enough to be noticeable in everyday contexts, such as when people perceive colors like turquoise differently. Holism would explain differences in turquoise perception in terms of different qualitative relations that place the relational node of turquoise closer to green–or blue, respectively. Therefore, interindividual differences are not an obstacle to holistic accounts; rather, they can be effectively explained.

A common question within structuralism is the extent to which holism should be applied. Structuralists might hold different views on this issue. Should a phenomenal quality be defined only within its specific quality space, or should we advocate a broader holism that encompasses multiple sensory modalities? Moreover, should holism extend to feature spaces of objects, potentially including object representations (Bellmund et al., 2018)? It seems plausible that structuralist definitions could apply across sensory domains. In fact, theories of iconic representations have suggested so in perceptual domains for decades (Burge, 2018). Another question is whether the same holds for cognitive domains. For example, the effectiveness of describing various cognitive domains, even language, in vector spaces suggests that structuralism could be extended to these domains (Bellmund et al., 2018). In this sense, structuralism could generalize key claims of well-known structural theories of representation in the domain of perception to a wider range of mental domains.

Different versions and degrees of structuralism remain possible. There is a prima facie logical independence between different forms of structuralism. Different structuralist positions can formulated depending on the mental domain they address. For example, structuralism regarding *perceptual* representations can be maintained independently of the view that *cognitive* representations also obey the structuralist principles of structural mapping, holistic individuation, and holistic exploitability. This could be called "horizontal" independence, since it relates to different structuralist perspectives within the same explanatory level, e.g. the representational level. Different structuralist views may differ in the mental domain they target.

In addition, "vertical" dissociations might be possible across different explanatory levels of cognition. Using Marr's (2010) framework of the computational, algorithmic, and implementational level, one could adopt a structuralist view at one level without requiring it at another. For example, a holistic representation at the computational level might not imply a physically holistic implementation. Similarly, phenomenality (either understood as computational or as a level in itself) could be holistic without requiring correlated algorithms or mechanisms to be holistic as well. But this would imply a very liberal reading of structuralism, rejecting the structural mirroring constraint.[2]

In my view, the label "structuralism" should be understood as an umbrella term linking a set of interrelated positions by the key hypotheses of structural individuation, structural mapping and structural exploitability.

In its most extreme form, an "overarching" version of structuralism would claim that all properties associated with mental functioning at each explanatory level and within each level are understood as relationally defined and holistically exploited. For example, structuralism about phenomenal experience could be combined with structuralism about mental representation, leading to a structuralist representationalism[3] about phe-

[2] If no mapping to the implementation level were assumed, one could speak of structuralism in the sense of structural individuation and exploitation. However, this would neither be a "neurophenomenal" (Fink et al., 2021) nor a "vehicle" structuralism (O'Brien & Opie, 2004).

[3] My colleague Daniel Weger has explored such a position in his dissertation, which he presented at the Models of Consciousness 5 conference.

nomenal experience (Fink & Kob, 2024). If this structural representationalism is conceptualized as vehicle representationalism (Lyre, 2022; O'Brien & Opie, 2004; O'Brien & Opie, 1999; Opie & O'Brien, 2015), this would lead to a view in which mental operations are hypothesized to amount to the three structuralist claims at *all* explanatory levels of the mind sciences. This overarching structuralism could even be extended to lower levels of the physical world if one additionally adopts the structural realism of the philosophy of science (Ladyman, 1998).

However, such an overarching view need not be adopted and seems costly to establish, since structuralist principles would have to be shown for each mental domain and it is unlikely that the mind uses one singular scheme for its representations, computations, algorithms, and implementations (as well as its phenomenality). Because to this indeterminacy in delineating structuralism as a specific position in the mind sciences, the following discussion of the relationship between structuralism and consciousness will be to some extent logically compatible with non-structuralist views of representation, neural implementation, and phenomenal properties. The advantage of this breadth is that the following discussions will also be relevant to people who are skeptical of the structuralist view.

Because of the explanatory gap (Levine, 1983), we cannot infer phenomenality directly from neural and behavioral data. This precludes to directly proof that phenomenality consists of structural properties only. My argument for structuralism is therefore pragmatic rather than metaphysical. I see the primary advantage of structuralism as its ability to provide a framework that integrates research and theory on neural implementation, representation, and phenomenality. Through its three core claims, structuralism provides a conceptual and methodological framework that links the technical languages of different fields and suggests new avenues for research. Notably, these core claims could hold even if there remains an intrinsic aspect to qualia—provided that qualia can be individuated by their relational properties, mapped onto neural structures, and structurally exploited. Thus, bypassing deeper metaphysical questions, I argue for structuralism on the grounds of its practical utility in interdisciplinary research. This paper thus seeks to establish methodological structuralism as a guiding principle in consciousness research.

3 Remarks on the notion of "content"

Before proceeding, some terminological clarification is necessary. The term "content," as I have used it goes beyond its traditional philosophical meaning. Typically, when discussing the consciousness-brain relationship independently of the mind-world relation, we consider phenomenal contents without reference to their connection to external states or affairs. However, philosophers generally define "content" in terms of its correspondence to states of the world. This raises a philosophical ambiguity: if "content" no longer refers to external events and affairs, it becomes unclear what a neural correlate of content should *represent*. Surely there is something odd about claiming that a neural correlate of experienced content represents phenomenal experience.

An alternative term to describe the neural correlate of phenomenal structure would be "neural correlate of conscious *character*" (Fink & Kob, 2024). With this terminology, the structuralist theory aims to provide a framework for understanding the neural encoding of the structure of subjective experience. However, unless one assumes that the qualitative properties of character can be fully explained solely in terms their structural properties alone (Fink et al., 2021; Loorits, 2014; Lyre, 2022; Tsuchiya & Saigo, 2021), the qualitative aspects themselves—such as the redness of red or the sensation of pain—may remain unexplained within the methodological structuralist view. In my understanding, while methodological structuralism allows researching the structure of phenomenal properties, because of the explanatory gap (Levine, 1983) it cannot address the qualitative nature of these experiences (see also Kleiner, 2024). Thus, the notion of "phenomenal character" does not fully serve our purposes here.

It follows that structuralism, as defined here, does not attempt to address the hard problem of consciousness (Chalmers, 1997). While it can be seen as a theory of how the structure of phenomenal character is neurally

encoded, it does not address the qualitative properties of consciousness—that are at the heart of the hard problem. A methodological structuralist perspective therefore does not aim to solve the hard problem and remains compatible with a range of metaphysical views, including dualistic approaches. In my view, methodological structuralism can be understood as an extended research program on Chalmers' (1997) "principle of structural coherence", which states that phenomenal properties and corresponding neural properties are structurally identical.

For lack of a better alternative, I will use the terms "conscious content" instead of "qualitative character", implying a loose meaning of "content." This is because the term "character" is rarely used in neuroscience, whereas "content" is widely used and often encompasses a broad range of meaning, including what philosophers would call "phenomenal character" (Marvan & Polák, 2020). Boly et al. (2023, p. 9604) for example write: "content-specific NCC are the neural mechanisms specifying particular phenomenal contents within consciousness, such as colors, faces, places, or thoughts." This usage makes no reference to the "aboutness" of these contents. Adopting this terminology may introduce some philosophical imprecision, for which I hope philosophical readers will be understanding.[4] I simply lack the space to thoroughly discuss philosophical questions about the relationship between phenomenal and representational properties (Kriegel, 2002; Pautz, 2010).

4 Limits and prospects of structuralism in the study of consciousness

The structuralist approach has several additional limitations, which are important for clarifying its role in the science of consciousness. First, we might consider the potential types of mappings: bijective (one-to-one in both directions), injective [one-to-one in one direction; Haynes (2009)] or surjective [many-to-one in one direction; Fink et al. (2021)]. If the consciousness-brain mapping were bijective, models of structures in subjective experience, such as color space, could directly reveal the corresponding neural structures by identifying neural systems that mirror the perceptual models. A bijective mapping implies a one-to-one correspondence, where each phenomenal color structure maps uniquely to a corresponding neural activation structure and vice versa. This would even allow for "reverse inference" (Poldrack, 2006), enabling inferences about consciously experienced content based on neural activation patterns alone.

It is unfortunate that bijectivity is highly implausible. Psychological states are generally considered to be multiply realizable, even within the same brain (Putnam, 1980). For instance, if a psychological state can be realized by two different neural systems, then the mapping from neural systems to psychological states would be inherently many-to-one. Consistent with classical accounts of supervenience (J. Kim, 1984) the set of potential neural states underlying consciousness appears to be larger than the set of possible conscious states. Consequently, it is more reasonable to assume a many-to-one surjective mapping, which can be expressed as a surjective homomorphism, or epimorphism (Fink et al., 2021).

Nevertheless, surjectivity imposes significant limitations on the structural constraint (Kleiner, 2024). It can only be used to narrow down potential neural correlates of phenomenal structure because, for any given structure in consciousness, multiple neural systems may exhibit a corresponding structure. Consequently, surjectivity serves as a limiting constraint on NCCs by identifying neural areas that exhibit the desired structure as potential candidates, while excluding others that do not. The structural constraint does not identify a single neural correlate of conscious content, but rather provides a set of candidate systems (Fink et al., 2021).

The structural constraint will yield a relatively large set of potential correlates of conscious content. This is because various neural systems are likely to exhibit content structures that resemble those found in subjective experience. For example, numerous neural systems, such as the lateral geniculate nucleus (LGN) and even retinal cones, exhibit color space structures similar to those in perceptual color space (Kuehni, 2003). However, it is not plausible to claim that conscious experience is encoded in the retina

[4] I would like to thank Andrew Lee for a very helpful conversation about my use of the term "content" at ASSC27 in Tokyo.

or LGN. As a result, while the structural constraint helps narrow down potential candidates, it cannot uniquely identify a specific neural correlate for a particular content domain.

The problem of multiple candidates can be understood as a matter of granularity. While retinotopic space and LGN space may exhibit structures similar to phenomenal color space, they are not structurally identical. Therefore, if a model of the specific structure of conscious color content were to be develop, the "grain problem" could be addressed. One approach would be to invent a highly detailed measure of the structures in subjective experience, assuming that the phenomenal color space has unique properties. For instance, if a maximally detailed description of all the sensory structures corresponding to the subject's immediate experience could be obtained, specific neural systems that encode this intricate structure could be identified with greater precision.

However, three significant problems arise (more limitations are discussed in Fink et al., 2021; Kleiner, 2024; Kob, 2023). *First*, there is a lack of a direct method for measuring the structures of subjective experience, necessitating the use of behavioral and physical proxies (Decock, 2006; Thompson et al., 1992).

Second, these proxies, such as behavioral outputs, may not accurately reflect the fine-grained structure of conscious experience. This discrepancy may be attributed to two potential factors: firstly, that behavior is dependent on neural pathways that differ from those involved in conscious awareness (Goodale & Milner, 1992); and secondly, that there is a loss of information between states of conscious experience and behavioral output. Consequently, in the absence of direct access to subjective experience, the structural constraint is limited in its ability to distinguish between multiple candidate systems as long as no direct measure of phenomenal structure is available. Nonetheless, structural mapping is a more restrictive and precise approach than mere correlation, making it a superior methodology to content encoding in traditional univariate approaches in the neuroscience of consciousness, which primarily focus on correlating consciousness with the statistical significance of neural activity in specific areas (Haynes, 2015).

Third, a highly significant issue concerns the relationship between the structure of perceptual models and phenomenal experience. Critics might argue that structural encoding concerns only the neural encoding of content in general, rather than *conscious* content in particular. In other words, structuralism might explain how content is neurally encoded, but not how it relates to conscious experience. Since the goal of consciousness science is to investigate conscious awareness rather than the general encoding of content, critics could contend that structuralists are misguided in associating their approach with consciousness science. This problem differs from the grain problem in that even a maximally detailed model of phenomenal structure might not be able to distinguish conscious from unconscious processing.

This is where the discussion becomes particularly relevant. I argue that the structuralist encoding model has the potential to uncover the neural basis of phenomenality itself (as opposed to unconscious perception), but only if there is something distinctive about the structure of conscious experience compared to unconscious perception (see also Fink & Kob, 2024; Fleming & Shea, 2024; Kob, 2023). The key question is the extent of the (in)dependence between neural content structures and subjective awareness. This is, in my view, the most promising avenue for neuroscientific investigation in a structuralist science of consciousness. Assuming a structural model of content encoding, progress can be made by exploring how the structure of neural encodings of perceptual content is related to the mechanisms underlying conscious awareness of these structures.

5 The one-factor/two-factor framework

Since the inception of modern consciousness science with Crick and Koch's work (Crick & Koch, 1990), the field has aimed to identify "neural factors" that differentiate conscious from unconscious perception (C.-Y. Kim & Blake, 2005). The underlying assumption is that conscious perception is neurally distinct from unconscious processing, and this distinction can be captured by contrasting neuroimaging data from both states (for review

and criticism see Aru et al., 2012; Irvine, 2013; Lepauvre & Melloni, 2021). These neural factors, which explain the differences between conscious and unconscious perception, are referred to as the neural correlates of consciousness (NCC). It is widely assumed that identifying NCCs provides a sufficient criterion for explaining conscious awareness in neuroscientific terms (Chalmers, 2000). According to this view, whenever the NCC mechanism is activated, phenomenal consciousness is present. However, many different neural factors could potentially account for consciousness, which limits the classical NCC approach's ability to differentiate between competing theories of consciousness (Fink, 2016; Paßler, 2023). Thus, the search for NCCs is compatible with a range of mechanisms that could exhibit consciousness (He, 2023; Zeki, 2003).

Consciousness science can progress with explicating different theories' assumption on the relationship between content and consciousness (Marvan & Polák, 2020). While some theories of consciousness propose that the mechanism responsible for general consciousness also explains the emergence of conscious content, many theories focus solely on distinguishing conscious from unconscious processing without addressing the neural encoding of content itself.

The latter type of theories aims to identify the neural basis for the distinction between conscious and unconscious states but does not address how content is encoded neurally. Nevertheless, such a theory must assume that the neural factor differentiating conscious from unconscious processing is somehow *linked* to content that is already encoded—either "making" it conscious or "leaving" it unconscious (for different views of this linkage, see Fleming & Shea, 2024). Consequently, despite the fact that these theories do not account for the neural encoding of content, they must explain how their proposed mechanisms of consciousness interact with the encoded content. Therefore, theories of consciousness can be classified based on whether they address both the content of experience and its conscious aspect or whether they explain consciousness in relation to content encoded by other neural systems and processes.

These approaches can be categorized as either one-factor or two-factor theories of consciousness (Fink & Kob, 2024; Marvan & Polák, 2020). The *one-factor view* posits that that a single mechanism underlies both consciousness and its content. In contrast, the *two-factor view* posits a division of labor: one factor explains the neural encoding of content, while another factor accounts for the mechanisms that make this content conscious. (See also Andrew Lee's (2024) very helpful discussion of the subtleties involved in the prominent metaphor of consciousness as a kind of inner light.)

From the two-factor perspective, a comprehensive explanation of conscious content would thus involve three components: *first*, identifying a neural mechanism responsible for consciousness; *second*, explaining the neural encoding of content; and *third*, modeling the interaction between the general consciousness-explaining factor and the content-explaining factor to clarify how content becomes conscious.

A two-factor view has a significant implication: it allows for the possibility that the structure of content can be identical in both conscious and unconscious processing. This is because, in this framework, content is explained independently of consciousness, suggesting that content and consciousness are separate or "orthogonal" (Vosgerau et al., 2008).

This assumption is likely to be readily accepted by many philosophers, as it is consistent with mainstream philosophical views. In philosophy, the notion of "content" is often defined as a representational relation between a mental system and external states of affairs. This implies that a neural system possesses content to the extent that it represents something external. Since these representations may or may not be part of conscious awareness, for many philosophers explaining these representations (and their structural properties) is a separate endeavor from explaining their occasional appearance in consciousness. Therefore, under the assumption of independence between content and consciousness, the properties of content, such as its structure, are independent of whether it is conscious or not. This implies that any kind of content structure can be potentially conscious or unconscious, which in turn implies the impossibility of identifying any "signatures" of consciousness in content structures.

An illustrative example of a two-factor theory is David Rosenthal's combination of higher-order thought theory with quality space theory

(e.g. Rosenthal, 2015, 2025). Rosenthal's approach exemplifies a "division of labor" between consciousness and qualitative content (or more precisely "character"). In accordance with this framework, the structural Quality Space Theory accounts for the qualitative character of perceptions, irrespective of whether they are conscious. In contrast, the Higher-Order Thought Theory explains how these perceptual qualities can enter subjective consciousness. A key implication of Rosenthal's theory is the requirement of equivalence between unconscious and conscious content: the same representations can be potentially processed consciously or unconsciously. Consequently, Higher-Order Thought Theory, when considered in conjunction with Quality Space Theory, serves as a prime exemplar of a two-factor theory of consciousness.

5.1 Most theories are two-factor theories

Most theories of consciousness are variants of the two-factor approach. For example, perceptual monitoring theory (Lau et al., 2022) proposes that content is encoded locally within sensory systems, and that awareness arises from the monitoring of this content. The Attention Schema Theory (Graziano, 2019; Graziano & Webb, 2015) posits that content becomes conscious when it is modeled by an attention schema that represents it. The Attended Intermediate Representation Theory (Prinz, 2012) explains content through the structure of so-called "vector waves" that encode intermediate-level representations, which become conscious via an attention mechanism (see also Marvan & Polák, 2020). Similarly, Global Neuronal Workspace Theory [GNWT; Dehaene and Changeux (2011); Mashour et al. (2020)] posits that while content is locally encoded by sensory systems, consciousness arises from its global distribution through a global workspace. Consequently, these theories exemplify various ways in which content and consciousness are related within a two-factor framework.

An interesting question pertains to the potential for two-factor theories to entail a re-representation of content in the neural system of the second factor (Fleming & Shea, 2024). This raises the question of whether a higher-order or GNWT system would need to replicate the structure of content from, say, a perceptual system. Such a "full content view" (Fleming & Shea, 2024, p. 7) would imply that the precise structure of subjective qualities is neurally computed by both local content and second-factor systems. However, "lean content" views (Fleming & Shea, 2024, p. 8), in which the second factor represents the structure of the content factor with a slightly modified structure, appear to be a more plausible explanation.

For example, the attention schema (Graziano, 2019; Graziano & Webb, 2015) would represent a simplified version of local content structures. However, if the Attention Schema Theory were the correct theory of consciousness, the structure of experience would be embedded in the attention schema system rather than in the local content system. Furthermore, as the attention schema is a simplified model, conscious experience would be less fine-grained in structure than local content representations. However, for the purposes of this discussion, it is sufficient to note that most theories employ two factors—a content explaining and a consciousness explaining factor—regardless of whether they entail a full or lean conception of content.

Crucially, not only "global" theories like GNWT can be interpreted as two-factor theories, but also Lamme's Recurrent Processing Theory (RPT) is consistent with this framework (Lamme, 2010, 2003, 2006; Lamme & Roelfsema, 2000). According to this theory, some contents can be encoded through both feedforward and recurrent processing, but only recurrent processing contributes to conscious experience. Other types of content, such as figure-ground segregation, even require recurrent processing and therefore only occur consciously in this view.[5]

However, this interpretation of the local recurrent processing view suggests that at least some content structures may remain equivalent across both feedforward and recurrent processing pathways. Nevertheless, it is the recurrent processing that is essential for making this content conscious. Thus, RPT can be interpreted as distinguishing between contents being processed, possibly feed-forward (one factor), and recurrent processing of these and other contents as another factor, which is the critical neural

[5] I thank an anonymous reviewer for pointing this out to me.

factor for explaining consciousness. In this view, while both feedforward and recurrent processes handle the same content, it is specifically the recurrent processing that is crucial for the content to achieve conscious awareness of this content.

5.2 One-factor theories

In contrast, a "one-factor view" posits that the process of making content conscious is integral to the construction of that content. Such views do not assume a strong division of labor between the neural encoding of perceptual structures and their conscious experience. Instead, they suggest that the structures encountered in conscious experience are intrinsically linked to the process of experiencing them. Examples of this perspective include first-order theories, which assert that the single contents present in sensory systems are conscious (Zeki & Bartels, 1999) and Integrated Information Theory (IIT), which links the structures of conscious experience directly to the mechanisms that produce consciousness (Albantakis et al., 2023; Tononi & Koch, 2015). Such a view can also be found in Chalmers' influential NCC article (Chalmers, 2000, p. 31): "an NCC (for content) is a minimal neural representational system N such that representation of a content in N is sufficient, under conditions C, for representation of that content in consciousness."

For instance, IIT posits that the structure of conceptual space, which influences the level of consciousness, also accounts for the structure of experience.[6] According to IIT, the structure of state transition probabilities in neural systems determines structures of conscious experience, such as the structure of color space (Tononi & Koch, 2015). In IIT, state space transition probabilities also affect Phi, the measure of integrated information. Hence, the conceptual structure determines the level of consciousness *as well as* the structure of experienced contents. Consequently, the structure of content, such as the experience of color, is directly linked to whether it is consciously experienced. If the structure of the state transition probability space were different, both the level of consciousness (Phi) *and* the structure of its contents would be different. Therefore, IIT explains both the emergence of specific content structures and the emergence of consciousness itself with a single theory, making it a one-factor theory.

It is important to clarify that one-factor theorists do not need to believe that certain contents can *only* occur consciously. To the contrary, the same content can potentially be conscious or unconscious on this view. When content occurs consciously, it is directly encoded by the neural states responsible for conscious awareness. In the one-factor view, however, content associated with the neural states constituting consciousness will be necessarily conscious, and content encoded by other neural systems necessarily unconscious. One-factor theories propose that, at any given moment, unconscious content is encoded by systems other than those that produce consciousness, while the conscious content experienced by the individual is encoded exclusively by the consciousness-producing systems.

6 Implications for AI consciousness

The structuralist theory of content can be plausibly applied to artificial systems like deep neural networks (DNNs) regardless of one's stance on consciousness. To the extent that DNNs exhibit stimulus-space structure in their activation spaces, they can be interpreted as processing content (Malach, 2021). The available evidence suggests this is indeed the case (Grossman et al., 2019; Kawakita et al., 2023; Marjieh et al., 2024), supporting the idea that DNNs encode content according to the structural approach. While a detailed discussion on the representational status of DNNs is beyond the scope of this paper, from a structuralist perspective, it seems at least plausible that they *could* be representational.[7] Thus, structuralists about

[6] I lack space for introducing IIT properly. For readers unaware of the theory, see Tononi and Koch (2015) for a nice introduction and Albantakis et al. (2023) for its most recent iteration.

[7] I am aware that I sidestep some important philosophical subtleties about to what extent structural similarity contributes to the processing of genuine representations independently of other, more general constraints on representations (Artiga, 2023; Shea, 2023). However, I hope the reader will forgive this, given the liberal notion of

mental content might hold the view that DNNs function as large-scale content-processing systems. If this is correct, then DNNs meet the criteria for a content factor. The question remaining is whether these contents could potentially become conscious.

In the context of a one-factor account, AI consciousness would depend on the intrinsic properties of the content-encoding system itself. Accordingly, solely from the perspective of a one-factor theory would today's DNN architectures, such as transformer-based large language models (LLMs), have to meet certain material or organizational criteria to potentially support consciousness. However, this line of thinking gives rise to an endless series of metaphysical debates concerning the nature of consciousness, whether it can be defined as a functional or material property (Kuhn, 2024; Seth, 2024).

Most theories of consciousness are two-factor theories and so questions of AI consciousness are best analyzed in a two-factor framework. From a structuralist perspective, DNNs may have sufficient properties to serve as a content factor. Accordingly, from a two-factor perspective, one of the two necessary factors for AI consciousness is satisfied. The crucial question now is whether the content encoded by DNNs could potentially become conscious if a suitable consciousness-conferring factor were to interact with it appropriately. Indeed, numerous AI labs are currently implementing this very approach, whereby a specific consciousness-conferring factor is introduced into the system, such as an attention schema, a global workspace, or recurrent processing (Butlin et al., 2023).

Note that the distinction between the bases of consciousness and its content allows for substrate-hybrid systems. That is, systems where content is processed by a different substrate than consciousness. If it turns out that consciousness needs a biological basis, perhaps because it requires integration at multiple scales (Seth, 2024), its contents could still be processed in a silicon-based system and the consciousness of those contents in an artificial organoid. The only requirement would be the construction of a suitable interface to transduce digital into neurochemical signals. Organoids (Pașca et al., 2024), neuroprostheses (Chen et al., 2020), and organoid-digital interfaces (Kagan et al., 2022) are active areas of research, so we can expect some progress here in the near future. For our purposes, it is sufficient to note that on the two-factor account, the question of AI consciousness is partly independent of the question of which kind of substrate is required to process content.

Admittedly, without an adequate way to measure consciousness, we will not be able to determine whether an artificial system is "truly" conscious, i.e., has subjective experiences as opposed to merely appearing so behaviorally. However, once the appropriate second factor is implemented, the system will likely exhibit behaviors indistinguishable from those of a conscious agent. This is because it will not only use the functional advantages that consciousness has in us, but will also possess the capacity to reflect its own states as either conscious or unconscious (Graziano, 2019). That is, AI systems could become something like the philosophical zombie (Chalmers, 1997), behaving like a conscious agent without any "inner lights turned on". For all practical purposes, these systems could become indistinguishable from conscious agents.

The important question for creating artificial conscious agents is to understand what kind of modulation of local content structure is done by consciousness. This will enable us to understand what kind of process consciousness is and what its functions are. In this way, methodological structuralism as a research program shifts from Chalmers' (1997) hard problem to Dennett's (2018, p. 1) hard question, "[O]nce some item or content 'enters consciousness', what does this cause or enable or modify?"[8]

content used in this paper, and the focus on the science of consciousness. Moreover, most of the additional constraints on representations can be well implemented in AI systems, so they do not raise any fundamental considerations against content processing in AI.

[8] Whether answering this question will enable us to create conscious agents in the phenomenal realist sense of the term "consciousness" or only in the sense of an illusionist interpretation of the term is ultimately beyond the scope of this paper.

7 Empirical considerations

The initial step for answering Dennett's (2018) hard question is to focus on differences in conscious and unconscious content structures. If a distinction in structure is identified, such that there are consciousness-specific structural properties in neural content, there is a high probability that insights can be gained regarding the functions of consciousness by understanding the alterations in content structure.

At first glance, one-factor theories suggest a difference between the structures of conscious and unconscious content, while two-factor theories imply equivalence (though the issue is more complex, as will be discussed shortly). In one-factor theories, the content structure and consciousness are interdependent, whereas in two-factor theories, they are not. Consequently, identifying differences in content structures for conscious and unconscious processing may facilitate the testing of theories against each other.

Interestingly, the (in-)equivalence of representations in conscious and unconscious processing has not been thoroughly tested. The one-factor/two-factor distinction is not merely a topic of philosophical debate; rather it has empirical implications within the structuralist framework. Structuralism can advance the study of consciousness by driving research that tests implications for theories of consciousness that arise from combining a proposed NCC with a structuralist view of content encoding.

The assumption of the independence of content structures from conscious awareness could theoretically be investigated by measuring content structures under conscious and unconscious experimental conditions and then comparing whether consciousness systematically influences content structures.

One approach involves using behavioral paradigms to compare conscious and unconscious color structures through metacontrast masking. In metacontrast masking, a target stimulus (such as a colored circle) is obscured by a mask (a ring with a hole matching the circle's size) that can be presented in different colors (Ro et al., 2009). When the circle is masked and presented unconsciously, its color can act as an unconscious prime for the visible ring. The premise is that if the unconscious color of the circle is similar to the conscious color of the ring, recognition of the ring's color should be facilitated. Reaction times are expected to be faster when the prime and the target colors are similar, even if the prime is processed unconsciously. If this facilitation effect is taken as a proxy for processing similarity of color content, a difference in the variances of this facilitation effect between conscious and unconscious processing could be cautiously interpreted as evidence that the similarity structure of unconscious processing is different from the structure of conscious color processing.[9]

Earlier studies have compared the structure of wavelength distributions emitted from a computer screen with priming responses under masked and unmasked conditions (Breitmeyer et al., 2004). These studies suggested that unconscious processing is more closely aligned with the physical wavelength structure, while conscious processing reflects psychological color similarity structures. The rationale is that unconscious colors are structured differently than conscious colors because unconscious processing is driven by the proximal stimulus, whereas conscious processing occurs at a later processing stage where illuminant and surface-related information are separated.

A subsequent study, however, contrasted surface color (after discounting the illuminant) with reflected color (including the illuminant) and indicated that early color processing might be structured according to surface color rather than the physical wavelength distributions reaching the eye (Norman et al., 2014). This suggests that unconscious color processing already entails psychological color similarities rather than reflecting the pure physical similarities of wavelength distributions. From this perspective, conscious and unconscious color appear to be similarly structured. Consequently, empirical science provides conflicting evidence in this regard.

[9] In fact, João Pedro Perreira Rodrigues, Karla Matić, Zefan Zheng, Marlo Paßler, Alex Lepauvre, Lucia Melloni, John-Dylan Haynes and I are currently conducting such a study. See also Zheng et al. (2024) for implications of unconscious priming studies for Global Neuronal Workspace Theory (Mashour et al., 2020).

Neuroscience may contribute to resolving the issue. A straightforward approach in the neuroscience of consciousness involves presenting both conscious and unconscious (e.g., masked) stimuli within a neuroimaging context and analyzing the resulting neural patterns. Representational similarity analysis (RSA) is particularly useful for this purpose (Kriegeskorte, 2008). RSA constructs representational dissimilarity matrices (RDMs), which map stimuli onto rows and columns, with the matrix' cells representing the dissimilarity between each pair of stimuli. To apply RSA, it is first necessary to compute dissimilarity measures for neural responses to different stimuli. These RDMs can then be compared to RDM's from various other datasets, such as psychological dissimilarities. By comparing these matrices, it is possible to evaluate the second-order similarity between different kinds of structures, such as psychological and neural structures.

The key step in this methodology in the context of consciousness science would be to correlate RDMs from conscious and unconscious conditions. This enables an estimation of whether the underlying content structures differ between conscious and unconscious processing. Specifically, by examining how the neural representation of stimuli changes from unconscious to conscious processing, it would be possible to infer whether and how content structures are modulated by consciousness.

Nevertheless, deriving meaningful pattern similarity structures for unconscious stimuli presents empirical challenges. One significant challenge is that the signal-to-noise ratio (SNR) for neural responses to unconscious stimuli can be insufficient, which complicates the computation of reliable distance measures between neural patterns (Mei et al., 2022). This issue can undermine the construction of accurate RDMs and limit the ability to compare neural activation structures effectively. Although RSA might be a fairly straightforward method to compare conscious and unconscious structure neurally, it might not be feasible for technical reasons.

Decoding studies also provide some insights into the relationship between neural representations and conscious versus unconscious processing. For example, Haynes and Rees (2005) demonstrated that neural classifiers can distinguish between different stimuli in binocular rivalry, indicating that a linear classifier can reliably differentiate neural data corresponding to one of the rivaling stimuli in conscious perception (but see Lin & He, 2009; Zou et al., 2016). Subsequent studies have expanded on these findings. Hesse and Tsao (2020) discovered that neurons in the inferotemporal cortex (IT) of macaques can dynamically switch between encoding conscious and unconscious stimuli, suggesting that neural activity related to conscious perception possesses distinct patterns.[10] Similarly, Sanchez et al. (2020) found that classifiers trained to distinguish between conscious and unconscious states in one modality could generalize to classify neural patterns from other modalities. This suggests that information about whether content is conscious is embedded in the code of the content encoding system (see also Graziano, 2019).

How do the decoding studies relate to the two-factor framework? At first glance, the supramodal activity signatures might seem to support the one-factor framework, indicating that content structures and their conscious experience may originate from a shared neural process. However, Sanchez et al. (2020), coming from a GNWT perspective, interpret the presence of common activity signatures across different sensory cortices as indicating a shared mechanism influencing these changes (but see He, 2023). This is consistent with what could be termed a *two-factor interaction view*, where consciousness and content mechanisms are distinct but interact with each other. In this perspective, a general factor that produces consciousness modulates the content-specific neural encodings in a specific manner, thereby rendering them conscious.

Nevertheless, decoding analyses themselves do not elucidate the structure of the neural patterns that they differentiate. From a geometric perspective, a decoder establishes a decision boundary between the neural patterns associated with different states it has been trained to recognize (Haynes, 2015). If a linear decision boundary can be drawn between clusters of patterns corresponding to various experimental conditions, the decoder will perform well. High decoding performance thus indicates that conscious

[10] This is not to say that the representations are conscious because of this activity. From the two-factor perspective, the IT cortex would explain the contents, and another factor would explain their consciousness.

and unconscious neural patterns are systematically mapped to distinct regions within a neural system's state space. Structuralism now encourages neuroscientists to further examine the pattern structures in these different state-space regions and to make predictions about the structural similarities or differences between conscious and unconscious neural patterns based on various theories of consciousness.

Investigating structural state-space differences between unconscious and conscious processing will enhance the understanding of the ways consciousness modulates local activity structure. There are several possible hypotheses regarding the relationship between decodability and neural activity structures.

First, decodability is consistent with both conscious and unconscious processing structures being equivalent. In general, the concept of decodability is coherent with the idea that conscious and unconscious patterns may exhibit a similar structural organization. In this case, consciousness could result in a systematic alteration in activity that affects all neurons within a system equally, thereby shifting the location of activity patterns within the system's state space without altering their structure.

A second possibility is that consciousness can be conceptualized as a *structuring* process, whereby neural patterns associated with unconscious processing are initially unstructured with respect to a particular stimulus space and become so structured only upon reaching conscious awareness.[11]

Third, consciousness could be closely associated with attentional processes. Attention is known to enhance clustering in the representations of local content, increasing both within-cluster similarity and between-cluster dissimilarity, thereby sharpening the contrast between representations (Brouwer & Heeger, 2013; Çukur et al., 2013). If consciousness functions as an attentional mechanism, we might expect not only a shift in the state space location but also enhanced clustering of neural patterns during conscious processing.

These are merely preliminary sketches of how models of shifts in state space structure between unconscious and conscious processing could serve as tests for theories of consciousness. Formulating precise hypotheses about structural (in)differences between conscious and unconscious processing of neural contents will eventually lead to new hypotheses along which the theories of consciousness can be pitted against each other. This is how I conceive of the advancement of consciousness science by the structuralist research program.

8 Limitations of the structuralist research program

Before concluding, it is important to address a general methodological caveat. Psychophysical techniques for inducing "invisibility" (C.-Y. Kim & Blake, 2005), such as masking and binocular rivalry, do not allow for a clean separation of consciousness from overall perceptual processing. These methods do not allow comparisons of the same representations in conscious and unconscious states; instead, they intervene at specific levels within the processing hierarchy (Breitmeyer, 2015), which introduces interactions between the experimental manipulation and the observed neural structures. As a result, comparisons between different methods may be confounded because each method perturbs processing at different levels in the processing hierarchy. Consequently, differences in the structure of representations associated with conscious versus unconscious processing may be partially attributable to the specific experimental techniques used. This suggests that structural similarities between conscious and unconscious processing are likely to be underestimated by the approach outlined in this paper.[12]

[11] As one reviewer pointed out, if this hypothesis is combined with a structuralist view of mental representation, there could be no unconscious representations. Although prima facie implausible from the perspective of cognitive science, philosophers have defended related claims in the context of phenomenal notions of intentionality (Kriegel, 2013).

[12] I thank an anonymous reviewer for suggesting a discussion of these issues.

Within the current paradigms of human psychophysics and neuroimaging, one way to addressing these issues lies in how we interpret studies that examine structural similarities between conscious and unconscious content. If there is a bias toward findings of structural differences, then finding structural similarity between conscious and unconscious processing to some degree could be taken evidence for the "same structure hypothesis", even if the similarity is not perfect. Conversely, if second-order correlations are consistently low or absent, this may be interpreted as evidence for the structural difference hypothesis. We can mitigate this methodological bias by adjusting our hypotheses accordingly.

We urgently need more theoretical frameworks to guide empirical research in consciousness science. This principle applies to all fields of scientific inquiry, but it is particularly important in the context of consciousness research, where the research target is so elusive. This paper aims to exploit the structuralist framework to improve our understanding of what to look for in neural data obtained from experimental paradigms that manipulate subjective awareness, both with current neuroimaging techniques and in light of future advances.

9 Conclusion

This paper uses structuralism as a framework to explore the relationship between the neural encoding of content and the neuroscience of conscious awareness. Adopting a two-factor view of conscious content – where one factor accounts for the neural encoding of content structure and the other for its appearance in subjective consciousness – I argue in favor of a research program that focuses on structural differences between conscious and unconscious processing. The structuralist methodology proves scientifically fruitful not in spite of, but because of, its precise assumptions about the neural encoding of content. As such, it represents a valuable approach for advancing the structural turn (Kleiner, 2024) in the neuroscience of consciousness.

Acknowledgments

I am grateful to two anonymous reviewers and to the editors of the journal. I would also like to thank the unconscious color research group, my colleagues in Magdeburg, and the students in my seminars for many stimulating discussions on structuralism and consciousness over the past few years.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms (L. J. Graham, Ed.). *PLOS Computational Biology*, *19*(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

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

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Bartels, A. (2005). *Strukturale Repräsentation*. Mentis.

Bartels, A. (2006). Defending the structural concept of representation. *THEORIA*, *21*(1), 7–19. https://doi.org/10.1387/theoria.550

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

Block, N. (2016). Semantics, conceptual role. In *Routledge encyclopedia of philosophy* (1st ed.). Routledge. https://doi.org/10.4324/9780415249126-W037-1

Bohon, K. S., Hermann, K. L., Hansen, T., & Conway, B. R. (2016). Representation of Perceptual Color Space in Macaque Posterior Inferior Temporal Cortex (the V4 Complex). *eneuro*, *3*(4), ENEURO.0039–16.2016. https://doi.org/10.1523/ENEURO.0039-16.2016

Breitmeyer, B. G. (2015). Psychophysical "blinding" methods reveal a functional hierarchy of unconscious visual processing. *Consciousness and Cognition*, *35*, 234–250. https://doi.org/10.1016/j.concog.2015.01.012

Breitmeyer, B. G., Ro, T., & Singhal, N. S. (2004). Unconscious Color Priming Occurs at Stimulus- Not Percept-Dependent Levels of Processing. *Psychological Science*, *15*(3), 198–202. https://doi.org/10.1111/j.0956-7976.2004.01503009.x

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *The Journal of Neuroscience*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Brouwer, G. J., & Heeger, D. J. (2013). Categorical Clustering of the Neural Representation of Color. *The Journal of Neuroscience*, *33*(39), 15454–15465. https://doi.org/10.1523/JNEUROSCI.2472-13.2013

Burge, T. (2018). Iconic Representation: Maps, Pictures, and Perception. In S. Wuppuluri & F. A. Doria (Eds.), *The map and the territory* (pp. 79–100). Springer International Publishing. https://doi.org/10.1007/978-3-319-72478-2_5

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness* (pp. 17–39). MIT Press.

Chalmers, D. J. (1997). *The conscious mind: in search of a fundamental theory* (1. issued as an Oxford University Press paperback). Oxford University Press.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. https://doi.org/10.1073/pnas.1402594111

Chen, X., Wang, F., Fernandez, E., & Roelfsema, P. R. (2020). Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. *Science*, *370*(6521), 1191–1196. https://doi.org/10.1126/science.abd7435

Churchland, P. S., & Sejnowski, T. J. (1999). *The computational brain* (5. print). MIT Press.

Churchland, P. M. (1986). Some Reductive Strategies in Cognitive Neurobiology. *Mind*, *95*(379), 279–309. https://doi.org/10.1093/mind/XCV.379.279

Churchland, P. M. (1993). State-space Semantics and Meaning Holism. *Philosophy and Phenomenological Research*, *53*(3), 667. https://doi.org/10.2307/2108090

Churchland, P. M., & Churchland, P. S. (1998). Conceptual Similarity across Sensory and Neural Diversity: The Fodor-Lepore Challenge Answered. In *On the contrary*. The MIT Press. https://doi.org/10.7551/mitpress/5123.003.0009

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., Van Den Bosch, J. J., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, *194*, 12–24. https://doi.org/10.1016/j.neuroimage.2019.03.031

CIE, C. I. d. l. ((1986). *Colorimetry* (2nd Edition, Vol. Publication CIE No. 15.2.). Commission Internationale de l'Eclairage.

Clark, A. (2000). *A theory of sentience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198238515.001.0001

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology*, *117*(1), 388–402. https://doi.org/10.1152/jn.00569.2016

Cohen, M. A., Nakayama, K., Konkle, T., Stantić, M., & Alvarez, G. A. (2015). Visual Awareness Is Limited by the Representational Architecture of the Visual System. *Journal of Cognitive Neuroscience*, *27*(11), 2240–2252. https://doi.org/10.1162/jocn_a_00855

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770. https://doi.org/10.1038/nn.3381

Cummins, R. (1995). *Meaning and mental representation* (3. print). 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

Davies, W. (2022). The paradox of colour constancy: Plotting the lower borders of perception. *Noûs*, *56*(4), 787–813. https://doi.org/10.1111/nous.12386

Decock, L. (2006). A physicalist reinterpretion of 'phenomenal' spaces. *Phenomenology and the Cognitive Sciences*, *5*(2), 197–225. https://doi.org/10.1007/s11097-005-9006-7

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170342. https://doi.org/10.1098/rstb.2017.0342

Diedrichsen, J., Yokoi, A., & Arbuckle, S. A. (2018). Pattern component modeling: A flexible approach for understanding the representational structure of brain

activity patterns. *NeuroImage*, *180*, 119–133. https://doi.org/10.1016/j.neuroimage.2017.08.051

Dretske, F. I. (1997). *Explaining behavior: reasons in a world of causes* (5th print). MIT Press.

Facchin, M. (2021). Structural representations do not meet the job description challenge. *Synthese*, *199*(3-4), 5479–5508. https://doi.org/10.1007/s11229-021-03032-8

Fink, S. B. (2016). A Deeper Look at the "Neural Correlate of Consciousness". *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01044

Fink, S. B., & Kob. (2024). Can structuralist theories be general theories of consciousness? In J. Hvorecký, T. Marvan, & M. Polák (Eds.), *Conscious and unconscious mentality: Examining their nature, similarities, and differences*. Routledge.

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*. https://doi.org/10.33735/phimisci.2021.79

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.2024.06.007

Fodor, J. A., & Lepore, E. (2004). *Holism: a shopper's guide* (Transf. to digital print). Blackwell.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, *32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Goodale, M. A., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Graziano, M. S. A. (2019). *Rethinking Consciousness*. W. W. Norton & Company, Incorporated.

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, 500. https://doi.org/10.3389/fpsyg.2015.00500

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., Khuvis, S., Herrero, J. L., Irani, M., Mehta, A. D., & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*(1), 4934. https://doi.org/10.1038/s41467-019-12623-6

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. https://doi.org/10.7554/eLife.56601

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals. *Trends in Cognitive Sciences*, *13*(5), 194–202. https://doi.org/10.1016/j.tics.2009.02.004

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Haynes, J.-D., & Rees, G. (2005). Predicting the Stream of Consciousness from Activity in Human Visual Cortex. *Current Biology*, *15*(14), 1301–1307. https://doi.org/10.1016/j.cub.2005.06.026

He, B. J. (2023). Towards a pluralistic neurobiological understanding of consciousness. *Trends in Cognitive Sciences*, *27*(5), 420–432. https://doi.org/10.1016/j.tics.2023.02.001

Hesse, J. K., & Tsao, D. Y. (2020). A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. *eLife*, *9*, e58360. https://doi.org/10.7554/eLife.58360

Irvine, E. (2013). *Consciousness as a scientific concept: A philosophy of science perspective*. Springer.

Isaac, A. M. C. (2013). Objective Similarity and Mental Representation. *Australasian Journal of Philosophy*, *91*(4), 683–704. https://doi.org/10.1080/00048402.2012.728233

Jazayeri, M., & Afraz, A. (2017). Navigating the Neural Space in Search of the Neural Code. *Neuron*, *93*(5), 1003–1014. https://doi.org/10.1016/j.neuron.2017.02.019

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kagan, B. J., Kitchen, A. C., Tran, N. T., Habibollahi, F., Khajehnejad, M., Parker, B. J., Bhat, A., Rollo, B., Razi, A., & Friston, K. J. (2022). In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron*, *110*(23), 3952–3969.e8. https://doi.org/10.1016/j.neuron.2022.09.001

Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., & Oizumi, M. (2023). Is my "red" your "red"?: Unsupervised alignment of qualia structures via optimal transport. https://doi.org/10.31234/osf.io/h3pqm

Kieval, P. H. (2022). Mapping representational mechanisms with deep neural networks. *Synthese*, *200*(3), 196. https://doi.org/10.1007/s11229-022-03694-y

Kim, C.-Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible'. *Trends in Cognitive Sciences*, *9*(8), 381–388. https://doi.org/10.1016/j.tics.2005.06.012

Kim, J. (1984). Concepts of Supervenience. *Philosophy and Phenomenological Research*, *45*(2), 153. https://doi.org/10.2307/2107423

Kleiner, J. (2024). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*, 103653. https://doi.org/10.1016/j.concog.2024.103653

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

Kriegel, U. (2002). Phenomenal content. *Erkenntnis*, *57*(2), 175–198. https://doi.org/10.1023/A:1020901206350

Kriegel, U. (Ed.). (2013). *Phenomenal intentionality*. Oxford University Press.

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

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*(1), 407–432. https://doi.org/10.1146/annurev-neuro-080317-061906

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

Kuehni, R. G. (2003). *Color Space and Its Divisions: Color Order from Antiquity to the Present* (1st ed.). Wiley. https://doi.org/10.1002/0471432261

Kuhn, R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, *190*, 28–169. https://doi.org/10.1016/j.pbiomolbio.2023.12.003

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, *13*(1), 47–76. https://doi.org/10.1080/09515080050002726

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

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220. https://doi.org/10.1080/17588921003731586

Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*(1), 12–18. https://doi.org/10.1016/S1364-6613(02)00013-X

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579. https://doi.org/10.1016/S0166-2236(00)01657-X

Lau, H., Michel, M., LeDoux, J. E., & Fleming, S. M. (2022). The mnemonic basis of subjective experience. *Nature Reviews Psychology*, *1*(8), 479–488. https://doi.org/10.1038/s44159-022-00068-6

Lee, A. Y. (2021). Modeling Mental Qualities. *The Philosophical Review*, *130*(2), 263–298. https://doi.org/10.1215/00318108-8809919

Lee, A. Y. (2024). The Light & the Room. In D. Curry & L. Daoust (Eds.), *Introducing philosophy of mind, today*. Routledge.

Lee, A. Y., Myers, J., & Rabin, G. O. (2023). The structure of analog representation. *Noûs*, *57*(1), 209–237. https://doi.org/10.1111/nous.12404

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

Lepauvre, A., & Melloni, L. (2021). The search for the neural correlate of consciousness: Progress and challenges. *Philosophy and the Mind Sciences*, *2*. https://doi.org/10.33735/phimisci.2021.87

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Lin, Z., & He, S. (2009). Seeing the invisible: The scope and limits of unconscious processing in binocular rivalry. *Progress in Neurobiology*, *87*(4), 195–211. https://doi.org/10.1016/j.pneurobio.2008.09.002

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

Lyre, H. (2022). Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1), niac012. https://doi.org/10.1093/nc/niac012

Malach, R. (2021). Local neuronal relational structures underlying the contents of human conscious experience. *Neuroscience of Consciousness*, *2021*(2), niab028. https://doi.org/10.1093/nc/niab028

Marjieh, R., Sucholutsky, I., Van Rijn, P., Jacoby, N., & Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, *14*(1), 21445. https://doi.org/10.1038/s41598-024-72071-1

Marr, D. (2010). *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press.

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*(2). https://doi.org/10.33735/phimisci.2020.II.61

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mei, N., Santana, R., & Soto, D. (2022). Assessing the brain representation of conscious and unconscious visual contents using encoding based representational similarity analysis. https://doi.org/10.1101/2022.12.23.521727

Morgan, A. (2014). Representations gone mental. *Synthese*, *191*(2), 213–244. https://doi.org/10.1007/s11229-013-0328-7

Morrison, J. (2020). Perceptual Variation and Structuralism. *Noûs*, *54*(2), 290–326. https://doi.org/10.1111/nous.12245

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435. https://doi.org/10.2307/2183914

Nirshberg, G., & Shapiro, L. (2021). Structural and indicator representations: A difference in degree, not kind. *Synthese*, *198*(8), 7647–7664. https://doi.org/10.1007/s11229-020-02537-y

Norman, L. J., Akins, K., Heywood, C. A., & Kentridge, R. W. (2014). Color Constancy for an Unseen Surface. *Current Biology*, *24*(23), 2822–2826. https://doi.org/10.1016/j.cub.2014.10.009

O'Brien, G., & Opie, J. (2004). Notes Toward a Structuralist Theory of Mental Representation. In *Representation in mind* (pp. 1–20). Elsevier. https://doi.org/10.1016/B978-008044394-2/50004-X

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*(1), 127–148. https://doi.org/10.1017/S0140525X9900179X

Opie, J. P., & O'Brien, G. J. (2015). The structure of phenomenal consciousness. In S. M. Miller (Ed.), *Advances in consciousness research* (pp. 445–464, Vol. 92). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.92.20opi

Pașca, S. P., Arlotta, P., Bateup, H. S., Camp, J. G., Cappello, S., Gage, F. H., Knoblich, J. A., Kriegstein, A. R., Lancaster, M. A., Ming, G.-L., Novarino, G., Okano, H., Parmar, M., Park, I.-H., Reiner, O., Song, H., Studer, L., Takahashi, J., Temple, S., … Young-Pearse, T. (2024). A framework for neural organoids, assembloids and transplantation studies. *Nature*. https://doi.org/10.1038/s41586-024-08487-6

Paßler, M. (2023). The exclusionary approach to consciousness. *Neuroscience of Consciousness*, *2023*(1), niad022. https://doi.org/10.1093/nc/niad022

Paßler, M., & Doerig, A. (2024). Neurophenomenal Structuralism and the Role of Computational Context. https://doi.org/10.48550/ARXIV.2412.20873

Pautz, A. (2010). Why Explain Visual Experience in Terms of Content? In B. Nanay (Ed.), *Perceiving the world* (pp. 254–309). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195386196.003.0010

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Prinz, J. J. (2012). *The conscious brain: how attention engenders experience*. Oxford university press.

Putnam, H. (1980). 17. The Nature of Mental States. In N. Block (Ed.), *The language and thought series*. Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c26

Ramsey, W. M. (2007). *Representation Reconsidered* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511597954

Ro, T., Singhal, N. S., Breitmeyer, B. G., & Garcia, J. O. (2009). Unconscious processing of color and form in metacontrast masking. *Perception & Psychophysics*, *71*(1), 95–103. https://doi.org/10.3758/APP.71.1.95

Rosenthal, D. (2015). Quality Spaces and Sensory Modalities. In P. Coates & S. Coleman (Eds.), *Phenomenal qualities* (pp. 33–65). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198712718.003.0002

Rosenthal, D. (2025). Objective foundations for the study of mental qualities. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11526

Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: Proxy vehicles and provisional representations. *Synthese*, *199*(3-4), 5917–5935. https://doi.org/10.1007/s11229-021-03052-4

Sanchez, G., Hartmann, T., Fuscà, M., Demarchi, G., & Weisz, N. (2020). Decoding across sensory modalities reveals common supramodal signatures of conscious perception. *Proceedings of the National Academy of Sciences*, *117*(13), 7437–7446. https://doi.org/10.1073/pnas.1912584117

Seth, A. (2024). Conscious artificial intelligence and biological naturalism. https://doi.org/10.31234/osf.io/tz6an

Shagrir, O. (2012). Structural Representations and the Brain. *The British Journal for the Philosophy of Science*, *63*(3), 519–545. https://doi.org/10.1093/bjps/axr038

Shea, N. (2018). *Representation in Cognitive Science* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198812883.001.0001

Shea, N. (2023). Organized representations forming a computationally useful processing structure. *Synthese*, *202*(6), 175. https://doi.org/10.1007/s11229-023-04373-2

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508. https://doi.org/10.1007/BF00499820

Thompson, E., Palacios, A., & Varela, F. J. (1992). Ways of coloring: Comparative color vision as a case study for cognitive science. *Behavioral and Brain Sciences*, *15*(1), 1–26. https://doi.org/10.1017/S0140525X00067248

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034. https://doi.org/10.1093/nc/niab034

Vosgerau, G., Schlicht, T., & Newen, A. (2008). Orthogonality of Phenomenality and Content. *American Philosophical Quarterly*, *45*(4), 309–328.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences*, *7*(5), 214–218. https://doi.org/10.1016/S1364-6613(03)00081-0

Zeki, S., & Bartels, A. (1999). Toward a Theory of Visual Consciousness. *Consciousness and Cognition*, *8*(2), 225–259. https://doi.org/10.1006/ccog.1999.0390

Zheng, Z.-F., Huang, S.-Y., Lu, S., & Cai, Y.-C. (2024). Interaction between top-down decision-driven congruency effect and bottom-up input-driven congruency effect is correlated with conscious awareness. *Journal of Experimental Psychology: General*, *153*(1), 102–121. https://doi.org/10.1037/xge0001483

Zou, J., He, S., & Zhang, P. (2016). Binocular rivalry from invisible patterns. *Proceedings of the National Academy of Sciences*, *113*(30), 8408–8413. https://doi.org/10.1073/pnas.1604816113