


Opinion

Quality space computations for consciousness

Stephen M. Fleming ^{1,2,3,4,7,9,*,@} and Nicholas Shea^{5,6,8,9,*}

The quality space hypothesis about conscious experience proposes that conscious sensory states are experienced in relation to other possible sensory states. For instance, the colour red is experienced as being more like orange, and less like green or blue. Recent empirical findings suggest that subjective similarity space can be explained in terms of similarities in neural activation patterns. Here, we consider how localist, workspace, and higher-order theories of consciousness can accommodate claims about the qualitative character of experience and functionally support a quality space. We review existing empirical evidence for each of these positions, and highlight novel experimental tools, such as altering local activation spaces via brain stimulation or behavioural training, that can distinguish these accounts.

Quality spaces and neural similarity structure

According to a long tradition in both psychology and philosophy, **conscious experiences** (see [Glossary](#)) form a similarity structure, namely a **quality space** [1–4]. Furthermore, the **quality space hypothesis** holds that relations between aspects of a conscious experience comprise part of its **phenomenal character** (i.e., of what it is like for the experiencing subject).

Empirically, **subjective similarity space** can be constructed directly from subjects' judgements, or can be measured indirectly, for example, using relative processing times [1,2,5] ([Box 1](#)). There is good evidence for the existence of quality spaces, at least within sensory modalities, some with multiple dimensions of similarity ([Box 2](#)). There are also several options for modelling quality spaces mathematically [4,6–8].

A core claim is that **subjective similarity** can be accounted for in terms of neural similarity, that is, in terms of the distance in neural activation space between patterns of neural activity responsible for realising experience. Methodological advances are providing increasingly rich data about such patterns, with techniques, such as **representational similarity analysis (RSA)**, being developed to analyse their similarity structure [9,10]. This allows us to identify correspondences between subjective similarity space and neural activation space. Empirical results suggest that neural similarity reflects subjective similarity [11–14] ([Figure 1](#)). For instance, different aspects of visual similarity, such as colour or shape, are tracked by activity patterns along the ventral temporal cortex [15–19], pointing to a distributed neural code for subjective similarity.

Theories of consciousness include more specific claims about the neural basis of conscious experience [20]. Some recent computational theories have achieved greater sophistication and testability by specifying, at least in outline, the computations that are responsible for conscious experience of individual stimuli (compared with processing such stimuli non-consciously). Here, we consider how such theories can account for the hypothesis that phenomenal character includes experiences of subjective similarity. For tractability, we focus on localist, global workspace, and **higher-order theories of consciousness**. These are sufficiently diverse to make distinct empirical predictions, and differ in how they accommodate existing evidence.

Highlights

The quality space hypothesis about conscious experience proposes that conscious sensory states are experienced in relation to other possible sensory states.

Recent empirical findings suggest that subjective quality space can be explained in terms of similarities in neural activation patterns; however, computational theories of consciousness have had little to say about this qualitative aspect of experience.

Localist, global workspace, and higher-order theories of consciousness can all accommodate a quality space in which subjects experience relations between actually instantiated experiences.

Different theories make distinct predictions for inactivation and stimulation experiments on both local (sensory) and nonlocal (e.g., prefrontal) neural activity patterns.

If experiments find that the whole structure of activation space affects the qualitative character of experience, this can be accommodated by localist and lean workspace/higher-order theories, but is more challenging for full-content global workspace and higher-order theories that require the whole structure to be re-represented.

¹Wellcome Centre for Human Neuroimaging, University College London, London, UK

²Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

³Department of Experimental Psychology, University College London, London, UK

⁴Canadian Institute for Advanced Research (CIFAR), Brain, Mind, and Consciousness Program, Toronto, ON, Canada

⁵Institute of Philosophy, School of Advanced Study, University of London, London, UK

Box 1. Probing psychological and neural similarity

Arbitrating between theories of the mechanistic implementation of subjective similarity spaces requires mapping between neural and psychological similarity [8,75]. Humans in empirical studies can easily and intuitively judge similarity between different stimuli, either with direct ratings of similarity, or within indirect tasks, such as choosing the odd one out (the least similar) of a triad, via graphically arranging similar stimuli together [78], or, in the triplet task, by choosing reference images that are most similar to a target [79].

The result of these empirical investigations is an $N \times N$ matrix, where each entry reflects the similarity between all pairs of the N stimuli under test. These data imply a psychological space, that is, how close one stimulus is to another in an arbitrary space. The idea is that what matters for defining a psychological space is not absolute quantities, but distances between different stimuli within a space. Similar matrices can be constructed from objective features of the stimuli, such as their image content or contrast, or from the outputs of machine learning methods, such as convolutional neural networks, applied to the stimulus set. Perceived similarity need not match objective similarity. For instance, for features obeying the Shepherd scaling law, perceived similarity falls off exponentially with distance in psychological space [3]. Distances can be computed within different data structures, such as metric distances in a geometric space or non-metric distances between nodes in a graph. A popular technique for inferring a geometric space from empirical data is multidimensional scaling (MDS), which attempts to find a low-dimensional representation of the distances between different stimuli.

A core hypothesis is that what matters for subjective similarity is the distance between two mental or neural representations [8]. Representational similarity analysis (RSA) is a commonly used technique that computes the pairwise similarity between neural patterns evoked by different stimuli, to create a representational dissimilarity matrix (RDM [9]). Neural RDMs can then be correlated with behavioural (or model) RDMs to identify correspondences between psychological and neural spaces. Again, what matters here is not the absolute properties of a neural representation, but the relative distance between one representation and another.

The 'structuralist' methodology is a way of applying these techniques to investigate the neural correlates of consciousness. It is promising because it characterises experiences in a more fine-grained way [75,80]. For example, while colours can be linearly decoded from activations in both V1 and V4, only the neural activity in V4 reflects the colour quality space, suggesting that activity in V4 and not V1 forms part of the basis of conscious colour experience [81,82].

⁶Faculty of Philosophy, University of Oxford, Oxford, UK

⁷<https://metacoglab.org/>

⁸www.nicholasshea.co.uk/

⁹These authors contributed equally to this work.

*Correspondence:

stephen.fleming@ucl.ac.uk

(S.M. Fleming) and

nicholas.shea@sas.ac.uk

(N. Shea).

[✉]X: [@smfleming](https://twitter.com/smfleming)

The standard assumption in consciousness science is that theories can 'mix and match' different accounts of the conscious–non-conscious contrast with different accounts of phenomenal character. We show that, in fact, different theories have more or less difficulty accommodating the quality space hypothesis. In particular, certain versions of workspace and higher-order theories have a problem with assimilating one version of the quality space hypothesis. Hence, perhaps surprisingly, the nature of phenomenal character furnishes evidence for and against some theories of consciousness.

How quality spaces relate to consciousness

Here we consider two ways in which the structure of quality spaces can be reflected in phenomenal character. A modest claim is that the properties that characterise conscious experience fall into families, such that properties within the same family (e.g., colour experiences) bear relations of phenomenal similarity and difference to one another. An analogy is the way in which the level of mercury in a thermometer represents temperature: different levels represent different temperatures, forming an 'organised' representational system [21]. The relational structure of temperature readings is latent in a given temperature reading, but is not represented by a single reading (Figure 2A). Under this view, the subject only experiences relations between aspects of their overall experiential state, that is, between aspects that they actually instantiate. There are both empirical and philosophical reasons for thinking that we experience colour relations, for example, in addition to categorical colour experiences [22–24]. We call this the 'instantiated relations' version of the quality space hypothesis.

A more arresting claim is that the phenomenal character of a single point in quality space, taken on its own, is affected by the overall structure of the quality space (Figure 2E). For instance, Lau argues that the phenomenal character of experience (i.e., 'what it's like' to have an experience) comprises experiencing how that experience is similar to other experiences ([25], pp. 197–198). Under this

Box 2. Characterising quality spaces and contrasting structuralism

Which types of conscious experience fall into quality spaces? It is generally accepted that sensory experiences do, not only the visual [12], but also auditory [83,84] and olfactory [85] experiences. Quality spaces could be restricted to perceptual or sensory experiences, but data on similarity spaces for more abstract categories [67] suggest that the phenomenon extends more widely, perhaps even characterising all conscious experiences. Psychological similarity spaces can be constructed for a range of categories, including facial identity, emotions, and the semantics of thoughts, although for each it is a further question whether the judged similarities are based only on similarity in phenomenal character.

Different kinds of experience appear to fall into distinct spaces. Where subjects find difficulty making consistent similarity judgements across categories, this suggests that the categories do not lie in a common subjective similarity space. For example, two stimuli may be similar in respect of colour and similar in respect of size, but the relative similarity of the colour of one to the size of the other may not be something that figures in experience. Further experiments, for example, on the consistency of similarity judgements within and between categories or sensory modalities, are needed to confirm this phenomenological intuition.

Experienced similarity requires some shared dimension or dimensions of similarity. For example, colour experiences are arranged into a quality space structured by three underlying dimensions: hue, brightness, and saturation. Experiences can be rated as similar separately along each of these dimensions. Even if experiences that share no common dimension, such as size and colour, are not experienced as bearing a relation of similarity/dissimilarity, humans may nonetheless be able to make judgements of similarity across experience types, for example between sound and shape [86].

The existence of quality spaces must be distinguished from structuralism about consciousness. Structuralism is a metaphysical claim about the nature of conscious experience. It is the claim that what makes a particular conscious experience *E* a conscious experience of a specific type is a matter of the relations of *E* to other experiences [87]. As a metaphysical claim, structuralism can be distinguished from structuralist methodology for studying consciousness (see Box 1 in the main text). The viability of structuralist methodology is independent of the truth or otherwise of structuralism as a metaphysical thesis [75]. For example, while the whole-structure version of the quality space hypothesis sits naturally with structuralism, even the instantiated relations version of the quality space hypothesis, or only the existence of quality spaces, allows us to use the similarity structure of experience to investigate its neural realization, even when those similarities are a contingent property and not part of what constitutes an experiential type. The strong claim of structuralism is not required to allow us to use similarity structure to infer that V4 but not V1 is responsible for realising colour experience (see Box 1 in the main text).

view, if the range over which experiences could occur across that space were different, even in a distant part of space, the subject's experience would be different. An analogy is the way in which the little blue dot on a smartphone map represents not only where the user is at that moment, but also the user's spatial relations to all the other points shown on the screen: if the map were larger or smaller, the content represented would be different. An upshot is that a creature that could only ever experience a single flash of light would not experience it as having any particular colour (cf. [25], p. 198). We call this the 'whole-structure' view.

In the whole-structure version of the quality space hypothesis, the overall structure of the quality space in which an experience *E* falls affects the phenomenal character of *E*. The instantiated relations view rejects this: only relations between experiences that are instantiated form part of phenomenal character. This still gives the subject some relational phenomenology (determined by the range of stimuli in play) and the impression that every aspect of experience has both a relational and a categorical phenomenal character (e.g., colour relations as well as intrinsic colour).

In the following sections, we consider how localist, global workspace, and higher-order theories of consciousness can accommodate these two distinct possibilities for a quality space. We consider how each combination of theory and quality space accommodates existing evidence, and makes distinct empirical predictions.

Quality spaces within localist theories of consciousness

Instantiated relations version of the quality space hypothesis

From the perspective of localist theories of consciousness [12,26–28], the varying activity patterns that support conscious experience are local to sensory systems (including, under local

Glossary

Conscious experience: mental state or event with a phenomenal character; thus, there is something it is like for a subject to have a conscious experience.

Full-content higher-order or workspace theory (of consciousness): sensory information only becomes conscious if it is appropriately recapitulated in other regions of the brain (in a higher-order representation or first-order re-representation/barcode).

Global neuronal workspace theory (GNWT): consciousness is constituted by the global broadcast of sensory representations through their being re-represented in a (prefrontal/parietal) central workspace.

Higher-order state space HOSS) model: consciousness is constituted by a higher-order mechanism that assigns a sufficiently high degree of reliability to first-order representations based on their magnitude or precision. First-order contents are gated into consciousness without being re-represented by the higher-order mechanism; a lean theory.

Higher-order theory (of consciousness): consciousness is constituted by higher-order representations of mental states.

Higher-order thought theory (of consciousness): consciousness is constituted by the content of thoughts about first-order (e.g., sensory) states; a full-content theory.

Lean higher-order or workspace theory (of consciousness): additional processing is needed for sensory information to become conscious, by being broadcast or reality tagged, but the sensory content need not be meta-represented or re-represented.

Perceptual reality monitoring (PRM): higher-order theory of consciousness in which sensory representations become conscious when indexed by a higher-order mechanism and categorised as reflecting external reality (rather than noise or imagination). PRM mechanisms are held to operate implicitly and continuously, and are not subject to volitional control.

Phenomenal character: subjective or experiential properties of a **conscious experience**; what it is like to undergo the experience (called 'subjective character' by some theorists).

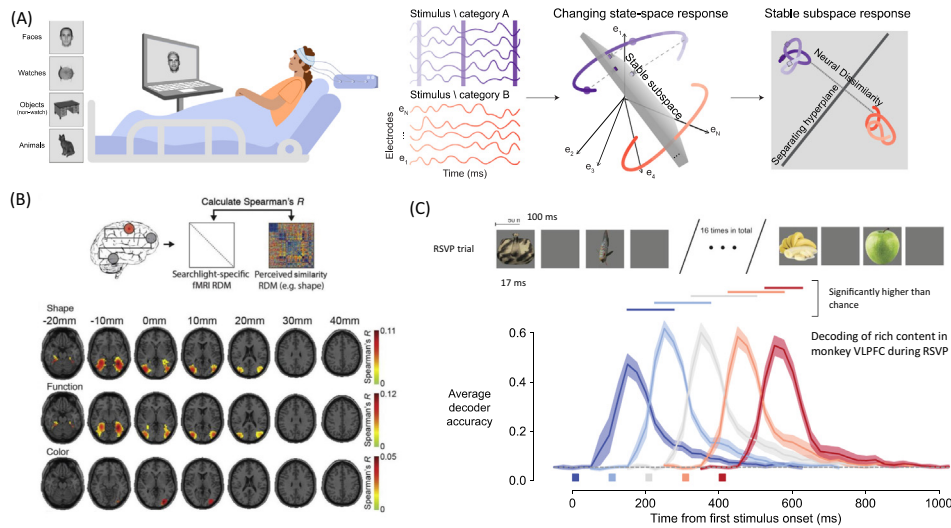


Figure 1. Quality spaces and neural similarity. (A) Intracranial recordings from patients viewing stimuli of variable duration from different categories [13]. The pattern of activation across recording sites in occipital and ventrotemporal cortex contains sustained and stable (invariant) information about the visual percept throughout the duration of the stimulus. Notably, while overall activity magnitude and firing pattern changes over time, there is a stable maintenance of neural dissimilarity between two stimulus category representations. (B) Location of representations underlying different facets of perceived visual similarity in the human brain [15]. A searchlight approach was used to create fMRI representational dissimilarity matrices (RDMs) in a sphere around each voxel, which were then correlated with perceived similarity RDMs for different features of a set of 118 object images (maps for shape, function, and color are shown; see [15] for other similarity dimensions). (C) Neuronal ensemble activity from the macaque ventrolateral prefrontal cortex (VLPFC) during briefly presented visual stimuli [51]. The identity of each stimulus could be robustly decoded from PFC population activity even under rapid serial visual presentation (RSVP) conditions, demonstrating a rich representation of visual content in prefrontal areas. Reproduced from [13] (A), [15] (B), and [51] (C).

recurrence theory, recurrent activity). Localist theories can readily accommodate the instantiated relations version of the quality space hypothesis.

One potential computational instantiation of this view is provided by population coding models of perception [29–31]. Under these models, different neural populations are tuned to different sensory attributes (e.g., motion direction or colour), and the combination of different tuned activation functions indexes a particular location in a (high-dimensional) perceptual space. The similarity between population codes is proposed to underpin psychological similarity [32]. Evidence for this view comes from recent studies mapping the psychological similarity of different visual stimulus dimensions (shape, colour, background, function, and total similarity) onto fMRI and magnetoencephalography (MEG) data. By partialing the contribution of each of several **representational dissimilarity matrixes (RDMs)**, different subregions within the ventral visual cortex were shown to correlate with different aspects of subjective similarity [15] (Figure 1B). Similar analyses can be performed on data from higher-resolution methods, such as single unit recording [33,34] and intracranial electroencephalography (IEEG) [11, 13,35,36], and RSA allows data from different resolutions (e.g., fMRI and single-unit recordings) to be compared in a common space [37].

One variant of the population-coding account holds that the brain is doing ‘linear algebra’ by combining activity along different (perhaps arbitrary) perceptual dimensions, with different levels of firing along different dimensions realising specific experiences. Studies of the neural encoding of faces in macaque inferotemporal (IT) cortex highlighted how individual faces are represented by

Qualitative character: phenomenal character, used to emphasise that conscious experiences or aspects thereof form a similarity structure (a quality space).

Quality space: subjective similarity space.

Quality space hypothesis: the phenomenal character of a conscious experience includes experiencing relations of subjective similarity between aspects of the experience (or between different experiences).

Representational dissimilarity: dissimilarity (metric or non-metric) in multivariate response space between two activity patterns (in a brain or computational model).

Representational dissimilarity matrix (RDM): an RDM for k stimuli is a $k \times k$ matrix of the representational dissimilarity in multivariate response space between all pairs of stimuli.

Representational geometry: arrangement in multivariate response space of a collection of the activity patterns produced in different conditions (e.g., in response to different stimuli); captured by an RDM.

Representational similarity analysis (RSA): statistical method that uses RDMs to compare the representational geometry of the responses produced in two spaces by a common set of stimuli; for example, in human versus monkey brains, brain data versus a computational model or, as here, as between subjective similarity space and neural activation space.

Subjective similarity: similarity in phenomenal character.

Subjective similarity space: multidimensional space (which need not be metric) in which conscious experiences or aspects thereof are arranged based on relative similarity/dissimilarity in their phenomenal character. Similarity is based on the phenomenal character of experiences themselves, not similarity of the objects or stimuli represented or presented in experience. A standard method uses subjects’ judgements about the latter to construct the former.

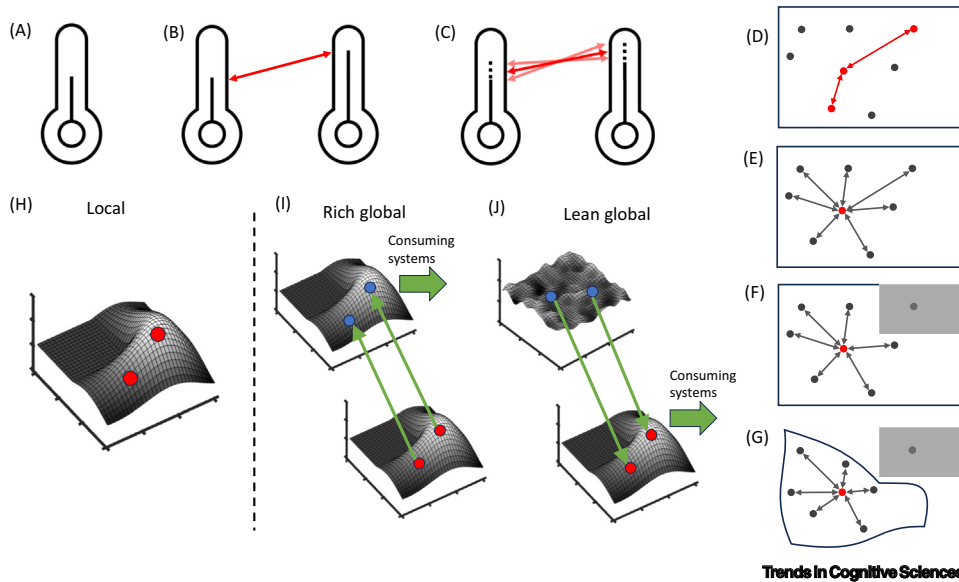


Figure 2. Theoretical options for capturing quality space. (A) The instantiated relations hypothesis about the nature of subjective quality space holds that experiences lie at a point along one or more dimensions, similar to the level of mercury in a thermometer. (B) Relations between currently active experiences (instantiated relations) are also experienced, but not the whole structure of the space. (C) Since each aspect of experience will in practice vary slightly over time, the subject will experience a whole family of such instantiated relations. (D) Switching analogy to points on a map, on the instantiated relations hypothesis, only relations between actually instantiated experiences are reflected in consciousness, indicated by the currently active red points on the map. (E) In the whole-structure hypothesis, the whole structure of the quality space affects the phenomenal character experienced at each point in the space, with the effect being constitutive not only causal. This is illustrated by the red point on the map, a currently experienced point in the space, having relational connections with other points in the space. (F) With a constitutive effect, deactivating or changing the activation space is like cutting out part of a paper sheet: the experience changes even though the local pattern of activation, the red point, stays the same. (G) A merely causal effect is like cutting out part of a rubber sheet: the experience changes because the space deforms, changing the focal activity pattern. (H–J) Illustration of how localist and global workspace/higher-order theories can accommodate quality spaces. Each subpanel illustrates points in high-dimensional neural activity space, here shown in three dimensions (e.g., the firing rate of three distinct neural populations). (H) Localist theories account for quality space in terms of location in a neural activation space. (I) According to full-content theories (higher-order or workspace), relational aspects are only experienced to the extent that those relations are recapitulated in the further states in which sensory information is re-represented or meta-represented. (J) According to lean theories (higher-order or playground), the relational content of sensory states forms part of experience in virtue of a global process without having to be recapitulated.

a linear combination of neural activations within a 50-dimensional ‘face space’ [38]. Notably, using binocular rivalry, both the consciously perceived and suppressed faces could be decoded from inferotemporal neuronal activations, but they occurred in distinct subspaces [39].

This view explains why local patterns of activation, which localist theories claim are the basis of specific contents becoming conscious, have a **qualitative character**. First, they exist in a similarity space, implemented in a neural activation space, the dimensions of which, as with size or colour, give the experience its particular character (just as a particular level of mercury in a thermometer falls in a similarity space with other levels of mercury). Second, relations between actually instantiated experiences (points in space) are themselves experienced. Since activation in response to a given input in practice changes over time, the subject will enjoy a range of different relations between experiences (Figure 2C). Thus, the actually instantiated relations may be rich.

An empirical prediction of the instantiated relations view is that interventions that alter the neural encoding space (e.g., preventing neurons encoding a particular dimension from firing) would not

change the phenomenal character of items that remain within the purview of the restricted neural space. One piece of evidence in support of this view is the finding that there are ‘null’ dimensions of neuronal manifolds along which changes in the stimulus do not affect neural responses [38,40].

Whole-structure version of the quality space hypothesis

There is also some empirical support for the whole-structure version of the quality space hypothesis, as treated by localist theories of consciousness. This view has been defended by Malach ([12], p. 7), who suggests that lateral cortical connectivity provides a neural substrate for the relational geometry. According to this account, local lateral connections in sensory brain areas implement a quality space for each sensory domain. Since **representational geometry** is defined by the neural activation patterns triggered by inputs, the brain network responsible will also depend on the strength of inputs to an area and on feedback connections. In turn, lateral connectivity itself (and, therefore, the quality space) will be modulated by top-down neuromodulatory connections.

Indirect evidence for such a view is provided by recent data showing that relative population activity patterns in visual cortex and their relational geometries are sustained during extended viewing of visual stimuli [11,13,35] (Figure 1A). Notably, while overall activation magnitude rapidly adapts, conscious perceptual content covaries with the multivariate neuronal pattern, which obeys constant similarity relationships with other stimulus-induced neuronal patterns. A notable feature of these neuronal patterns in posterior visual areas is that the representational geometry remains stable and unchanging over time, unlike the findings from prefrontal areas, which show more transient responses [13]. Thus, according to these data, what appears to matter for the experience of specific perceptual content is the geometry in which neural patterns are arranged, rather than absolute features of activation magnitude (Box 1). Intriguingly, similar representational geometry is observed in the early layers of deep neural networks trained on visual tasks [41]. However, such data could be (computationally) explained with the kind of population-coding dynamics described in the previous section, given that the same relational geometry can be instantiated with different neuronal implementations [42].

It might appear strange that relations to merely possible experiences should affect phenomenal character, but this view makes that un-mysterious, since the wider brain network, which is actual, not just potential, is a realiser for this (extremely rich) phenomenal character. An analogy is the way in which a glowing blue dot on a map represents where I am now. It represents my relations to all the other points on the map. The variable element is the blue dot, but the background map, which stays fixed, is a realiser for a very rich relational content.

A key prediction of the whole-structure view of qualitative character (unlike the instantiated relations view) is that, if one intervenes to change the structure of the activation space, and then manages to re-instantiate the same local pattern (e.g., using optogenetic reactivation methods), then what is experienced by the subject should differ even though the distributed pattern of activation remains the same. Testing this hypothesis requires more than just a causal change to the pattern of activation produced in response to input, for example following adaptation or learning. The claim here is that the whole space makes a constitutive difference: what is experienced will change even if the space is not deformed (similar to cutting out part of a paper map; Figure 2F). Some intriguing psychophysical evidence for this view was obtained by Song *et al.* [43]. By using a Hebbian training protocol to strengthen lateral connections between two points of visual space, the perceived distance between untrained locations was also altered. These findings suggest that a local change in the strength of lateral connections is sufficient to alter experience of visuospatial relations.

Quality spaces within workspace and higher-order theories of consciousness

We now turn to global workspace and higher-order theories of consciousness. Workspace accounts, such as the **global neuronal workspace theory (GNWT)**, posit that what is required is the broadcast or sharing of sensory information into a central workspace [44–46]. Higher-order theories propose that what is important for consciousness is the re-representation or monitoring of first-order sensory information [47–49].

David Rosenthal's **higher-order thought theory** holds that phenomenal character is fixed by the content of a higher-order thought that represents a perceptual representation [2]. Stanislas Dehaene's version of GNWT calls for sensory information to be re-represented and broadcast to become conscious [44,46]. The broadcast representations are supposed to act like a 'barcode' that encodes the broadcast content in full. Unlike higher-order theories, the broadcast representations are first-order. However, both theories are committed to phenomenal character depending on the content being carried in full by the higher-order or broadcast representation: they share a commitment to a **full-content** view of what is required for a content to be conscious. Since this commitment generates shared issues for accommodating the quality space hypothesis, we discuss these two theories in the same section, despite the important differences between them. Other higher-order and workspace theories require less of the second component, as we shall see. We discuss these **lean** accounts in the following section.

Full-content higher-order and workspace theories

According to Rosenthal's **higher-order theory**, qualitative character is given by higher-order thoughts that represent mental qualities in terms of locations in a quality space ([2], p. 383). He also holds that the subject can appreciate the relation between two mental qualities that occur concurrently and, thus, discriminate between them, but not between experiences that occur at different times ([2], p. 383). Instantiated relations contribute to phenomenal character, but the whole structure of quality space does not.

Dehaene's version of GNWT also subscribes to a full-content view of what is required for a content to become conscious. Evidence for a full re-representation of perceptual content in the global workspace comes from recent intracranial recording studies in both humans [13,35] and non-human primates [50–52]. These studies used either passive viewing or binocular rivalry to reveal the coding of multiple visual dimensions in prefrontal cortex (PFC), even when subjective reports were not required [50] (Figure 1C). However, it remains unclear how such decoding of contents relates to the representational spaces or population codes that have been identified in sensory areas (e.g., the face–space population code in IT cortex discussed earlier) or whether it recapitulates the RDM of perceptual similarity judgements. Compared with sustained representations in sensory areas, the PFC shows more transient coding of stimulus content at stimulus onset [13,35], together with the emergence of content-invariant visibility or 'phenomenal magnitude' signals [53–55].

Thus, it is unclear whether and how the PFC instantiates a quality space, as opposed to 'reading out' task-dependent aspects of organised, localist representations. One perspective on this question is provided by studies of the task dependence of PFC representations. Unlike in sensory areas, the dimensions of the neural code representing particular sensory features can be radically compressed or elongated depending on whether those features are task relevant [56–58]. Such a flexible reformatting of sensory representations in PFC suggests that global representations are less tied to the phenomenal character of conscious experience. However, whether phenomenal character is also unchanged in such scenarios remains an empirical question, given there is evidence for attention- and task-dependent changes in phenomenal character as measured by similarity judgments [4,59].

A theory intermediate between the full-content and lean views is Lau's **perceptual reality monitoring (PRM)** variant of higher-order theory [60]. PRM starts from the observation that the mere presence of a sensory signal does not create conscious experience; instead, consciousness ensues when such signals are categorised as reflecting external reality. PRM proposes that this difference is underpinned by a downstream higher-order mechanism that indexes and interprets the first-order representations, akin to a discriminator in a generative adversarial network (GAN [61]). Extending PRM to encompass quality spaces, Lau ([25], pp. 197–211) and Lau *et al.* [62] propose that a higher-order monitoring mechanism has access to the nature of neuronal coding in sensory cortices, similar to how feature-specific attention targets specific neuronal populations [63,64]. The similarity between higher-order indices encodes implicit knowledge of the structure of sensory spaces. It is the similarity between indices, not the similarity between the first-order states indexed, that gives rise to phenomenal character. PRM readily accounts for experienced similarity between actually instantiated experiences, to the extent that their relations are recapitulated in PFC [25,62], but faces more difficulty accommodating the whole-structure view.

A prediction of both full-content views is that aspects of the similarity structure embodied in local/first-order neuronal representations that are not structural features of the higher-order or broadcast state will not contribute to phenomenal character, even though they may affect behaviour. This prediction is difficult to test, as it requires interfering simultaneously on putative local (sensory) and global neuronal representations, but may be possible with optogenetic manipulation approaches, or the use of behavioural training approaches in combination with brain imaging. Indirect evidence may be obtained by asking how noninvasive brain stimulation [e.g., transcranial magnetic stimulation (TMS)] or lesions to PFC affect the structure of perceptual similarity judgments. Previous studies demonstrated the recruitment of abstract grid and place codes in PFC and hippocampus when navigating olfactory [65] and auditory [66] perceptual spaces. Such neuronal codes may support implicit access to relations between different points in quality space, as posited by PRM. However, the mapping between grid codes and conscious experience remains to be determined. One concern is that, if a similarity space is explicitly learned, these abstract codes may reflect the requirements of the navigation task, rather than reflecting perceptual experience [67].

Lean higher-order and workspace theories

A leaner version of higher-order theory is provided by the **higher-order state space (HOSS)** account, which proposes that a higher-order state 'tags' first-order perceptual content with a degree of reliability or experiential magnitude [68]. Contents become conscious when the level of first-order magnitude or precision crosses a certain threshold. Under this view, the contents of perception are determined by first-order activations within sensory systems (similar to localist accounts), and no aspect needs to be re-represented to form part of conscious experience. Instead, the higher-order state acts to 'gate' particular contents into conscious experience when they are considered reliable enough [69,70].

The contribution of a magnitude signal to reality monitoring has gained recent empirical support through observations that, in near-threshold perception, signals from perception and imagination jointly determine reality monitoring judgments [71]. However, a lean higher-order theory in which PFC and parietal cortex only support higher-order (and not also first-order) states would not predict findings that transient activations in these regions also show rich relational representations of perceptual content [13,35,70] and that prediction errors on perceptual magnitude in medial PFC also encode features of perceptual content [72] (Figure 1C).

It is straightforward for the HOSS model to account for the instantiated relations version of the quality space hypothesis. It can also readily accommodate the whole-structure hypothesis,

since it does not call for recapitulation of first-order contents, unlike PRM, which requires the whole structure of quality space to be recapitulated in similarities between higher-order indices. If the experience E of a point in quality space is indeed affected by the whole structure of the space, then an activity pattern-plus-brain-network is a good candidate to realise that experience, with a higher-order gating mechanism being responsible for the whole of that first-order content becoming conscious.

Lean workspace accounts also hold that, while an additional process is needed to make sensory representations globally available, hence conscious, the first-order contents need not be re-represented. One example is the 'global playground': a broad network of areas among which sensory representations are shared and maintained for several hundreds of milliseconds, thus offering wider cognitive possibilities compared with automatic unconscious processing, but with no specific agenda [73]. The global playground does not require contents to be re-represented to become conscious. Nevertheless, representations in the global playground have a different functional signature compared with non-conscious representations.

Similar to localist theories, and like the HOSS model, the global playground can readily account for conscious experience of relations of subjective similarity. When a representation in a similarity space becomes conscious, the experiential property exists in a space of possibilities, in analogy to the range of levels of mercury in a thermometer. The nature of the experience is determined by its location in that quality space, which is realised by the location of a pattern of neural activity in neural activation space. Furthermore, relations between aspects of experience are realised by relations between occurrent neural patterns and, thus, also form part of the phenomenal character of experience, in line with the instantiated relations view. Lean workspace theories can also accommodate the whole-structure view, in the same way as lean higher-order theories, since no re-representation is required.

A prediction of lean higher-order accounts is that, if the relations between stimulus features remain stable (correlated), even in the face of absolute changes in activation, such a configuration will be more likely to pass a gating mechanism and, thus, become conscious. Features that are each too noisy to become conscious might, if the noise is correlated, be gated into consciousness together. Some preliminary evidence for this hypothesis comes from experiments showing that more stable neural activity patterns are associated with conscious perception [74]. A prediction of the global playground view is that conscious representations have a similarity structure that is different from, and richer than, that instantiated by non-conscious representations [39,75–77] (Box 1).

Concluding remarks

We have considered how neural similarity structures can be integrated into current computational accounts of consciousness. Localist, higher-order, and workspace theories can all account for the fact that experiences occur in a quality space. They can also readily accommodate the hypothesis that subjects experience relations between actually instantiated aspects of experience. If empirical investigations find, more arrestingly, that the whole structure of quality space affects the phenomenal character of each individual experience, that finding can be accommodated by localist theories, and also by lean higher-order and lean workspace theories. It is more problematic for full-content higher-order and full-content workspace theories, since they would require the whole structure to be re-represented.

Notably these views make distinct predictions for inactivation and stimulation experiments on both local (sensory) and global (e.g., prefrontal) neural activity patterns. These techniques are currently most accessible in animal models. While consciously experienced qualities are difficult to

Outstanding questions

Are properties of stimuli that are represented non-consciously (e.g., chromatic stimulation as opposed to perceived colour) represented in similarity spaces or in some other way? If in similarity spaces, are there systematic differences between conscious and non-conscious similarity spaces? More ambitiously, is there a characteristic type of similarity space that goes along with states being conscious?

Is experience of actually instantiated relations between experiences restricted to those occurring at a moment, or does it extend to relations between aspects of experiences occurring at nearby times?

Does noise/stability in relations in sensory content affect whether those contents become conscious (in line with a quality space version of lean higher-order theories of consciousness)?

Can behavioural adaptation approaches be extended to ask about the role of adaptation of local connectivity in other kinds of sensory experience (both visual and non-visual)?

How can higher-order index views be tested at the level of neural implementation? What would a neural signature of an index and its associated similarity space look like?

Must similarities be recapitulated in higher-order indices to be experienced (as claimed by PRM), or is it sufficient that first-order similarity structure is tagged as real and thereby gated into experience (as claimed by HOSS)?

How do quality spaces depend on the temporal structure of consciousness, and the potential for extended windows of integration? How do quality spaces change over time, given that neural similarity structures are themselves distributed in space and time?

How are relations encoded for simultaneously presented features (as anticipated under the instantiated relations view)?

probe in animals, a first step would be to combine inactivation methods with behavioural techniques for probing psychological similarity. Our overview also provides an organising framework for interpreting correlational data from studies comparing neural similarity structures obtained for conscious versus non-conscious representations in the human brain (see [Outstanding questions](#)).

Computational accounts of conscious experience to date have had little to say about the relational aspects of phenomenal character. We hope that this survey sets an agenda for incorporating the quality space hypothesis into existing theories of consciousness, at a time when emerging new data in the neurosciences make addressing this question both pressing and tractable.

Acknowledgements

S.M.F. is a CIFAR Fellow in the Brain, Mind, & Consciousness Program, and funded by UK Research and Innovation (UKRI) under the UK Government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society.

Declaration of interests

The authors declare no competing interests.

References

- Clark, A. (1996) *Sensory Qualities (revised)*, Oxford University Press
- Rosenthal, D. (2010) How to think about mental qualities. *Philos. Issues* 20, 368–393
- Shepard, R.N. (1987) Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323
- Nosofsky, R.M. (1992) Similarity scaling and cognitive process models. *Annu. Rev. Psychol.* 43, 25–53
- Cohen, M.A. *et al.* (2015) Visual awareness is limited by the representational architecture of the visual system. *J. Cogn. Neurosci.* 27, 2240–2252
- Lee, A.Y. (2021) Modeling mental qualities. *Philos. Rev.* 130, 263–298
- Tsuchiya, N. and Saigo, H. (2021) A relational approach to consciousness: categories of level and contents of consciousness. *Neurosci. Conscious.* 2021, niab034
- Roads, B.D. and Love, B.C. (2023) Modeling similarity and psychological space. *Annu. Rev. Psychol.* 75, 215–240
- Kriegeskorte, N. and Kievit, R.A. (2013) Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17, 401–412
- Robinson, A.K. *et al.* (2023) Visual representations: insights from neural decoding. *Annu. Rev. Vis. Sci.* 9, 313–335
- Broday-Dvir, R. *et al.* (2023) Perceptual stability reflected in neuronal pattern similarities in human visual cortex. *Cell Rep.* 42, 112614
- Malach, R. (2021) Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci. Conscious.* 2021, niab028
- Vishne, G. *et al.* (2023) Distinct ventral stream and prefrontal cortex representational dynamics during sustained conscious visual perception. *Cell Rep.* 42, 112752
- Rosenthal, I.A. *et al.* (2021) Color space geometry uncovered with magnetoencephalography. *Curr. Biol.* 31, 515–526
- Cichy, R.M. *et al.* (2019) The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage* 194, 12–24
- Op de Beeck, H.P. *et al.* (2008) Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci.* 28, 10111–10123
- Grill-Spector, K. and Weiner, K.S. (2014) The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15, 536–548
- Ferko, K.M. *et al.* (2022) Activity in perirhinal and entorhinal cortex predicts perceived visual similarities among category exemplars with highest precision. *eLife* 11, e66884
- Bracci, S. *et al.* (2019) The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *J. Neurosci.* 39, 6513–6525
- Seth, A.K. and Bayne, T. (2022) Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452
- Shea, N. (2023) Organized representations forming a computationally useful processing structure. *Synthese* 202, 175
- Papineau, D. (2015) Can we really see a million colours? In *Phenomenal Qualities: Sense, Perception, and Consciousness* (Coates, P. and Coleman, S., eds), pp. 274–297, Oxford University Press
- Morrison, J. (2020) Perceptual variation and structuralism. *Noûs* 54, 290–326
- Davies, W. (2021) Colour relations in form. *Philos. Phenomenol. Res.* 102, 574–594
- Lau, H. (2022) *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*, Oxford University Press
- Lamme, V.A.F. (2006) Towards a true neural stance on consciousness. *Trends Cogn. Sci.* 10, 494–501
- Block, N. (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–499
- Zeki, S. and Bartels, A. (1998) The autonomy of the visual systems and the modularity of conscious vision. *Philos. Trans. R. Soc. B Biol. Sci.* 353, 1911–1914
- Panzeri, S. *et al.* (2015) Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* 19, 162–172
- Barlow, H.B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 371–394
- Averbeck, B.B. *et al.* (2006) Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366
- Kriegeskorte, N. (2009) Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* 3, 363–373
- Reber, T.P. *et al.* (2019) Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. *PLoS Biol.* 17, e3000290
- Kriegeskorte, N. *et al.* (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141
- Cogitate Consortium *et al.* (2023) An adversarial collaboration to critically evaluate theories of consciousness. *bioRxiv*, Published

- online June 26, 2023. <https://doi.org/10.1101/2023.06.23.546249>
36. Davidesco, I. *et al.* (2014) Exemplar selectivity reflects perceptual similarities in the human fusiform cortex. *Cereb. Cortex* 24, 1879–1893
 37. Barron, H.C. *et al.* (2021) Cross-species neuroscience: closing the explanatory gap. *Philos. Trans. R. Soc. B Biol. Sci.* 376, 20190633
 38. Chang, L. and Tsao, D.Y. (2017) The code for facial identity in the primate brain. *Cell* 169, 1013–1028
 39. Hesse, J.K. and Tsao, D.Y. (2020) A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. *eLife* 9, e58360
 40. Bao, P. *et al.* (2020) A map of object space in primate inferotemporal cortex. *Nature* 583, 103–108
 41. Yamins, D.L.K. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624
 42. Kriegeskorte, N. and Wei, X.-X. (2021) Neural tuning and representational geometry. *Nat. Rev. Neurosci.* 22, 703–718
 43. Song, C. *et al.* (2017) Plasticity in the structure of visual space. *eNeuro* 4, ENEURO.0080-17.2017
 44. Mashour, G.A. *et al.* (2020) Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798
 45. Dehaene, S. *et al.* (1998) A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. USA* 95, 14529–14534
 46. Dehaene, S. *et al.* (2014) Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* 25, 76–84
 47. Rosenthal, D.M. (2005) *Consciousness and Mind*, Oxford University Press
 48. Lau, H.C. and Rosenthal, D. (2011) Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373
 49. Brown, R. *et al.* (2019) Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768
 50. Panagiotaropoulos, T.I. *et al.* (2020) Prefrontal cortex and consciousness: beware of the signals. *Trends Cogn. Sci.* 24, 343–344
 51. Bellet, J. *et al.* (2022) Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neurosci. Conscious.* 2022, niac005
 52. Kapoor, V. *et al.* (2022) Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat. Commun.* 13, 1535
 53. Podvalny, E. *et al.* (2019) A dual role of prestimulus spontaneous neural activity in visual object recognition. *Nat. Commun.* 10, 3910
 54. Barnett, B. *et al.* (2024) Identifying content-invariant neural signatures of perceptual vividness. *PNAS Nexus* 3, pgae061
 55. Hatamimajoumerd, E. *et al.* (2022) Decoding perceptual awareness across the brain with a no-report fMRI masking paradigm. *Curr. Biol.* 32, 4139–4149
 56. Mack, M.L. *et al.* (2020) Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* 11, 46
 57. Flesch, T. *et al.* (2022) Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* 110, 1258–1270
 58. Castegnetti, G. *et al.* (2021) How usefulness shapes neural representations during goal-directed behavior. *Sci Adv* 7, eabd5363
 59. Carrasco, M. *et al.* (2004) Attention alters appearance. *Nat. Neurosci.* 7, 308–313
 60. Lau, H. (2019) Consciousness, metacognition, & perceptual reality monitoring. *PsyArXiv*, Published online June 10, 2019. <https://dx.doi.org/10.31234/osf.io/ckbyf>
 61. Gershman, S.J. (2019) The generative adversarial brain. *Front. Artif. Intell.* 2, 18
 62. Lau, H. *et al.* (2022) The mnemonic basis of subjective experience. *Nat. Rev. Psychol.* 1, 479–488
 63. Bichot, N.P. *et al.* (2019) The role of prefrontal cortex in the control of feature attention in area V4. *Nat. Commun.* 10, 5727
 64. Goddard, E. *et al.* (2022) Spatial and feature-selective attention have distinct, interacting effects on population-level tuning. *J. Cogn. Neurosci.* 34, 290–312
 65. Bao, X. *et al.* (2019) Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron* 102, 1066–1075
 66. Aronov, D. *et al.* (2017) Mapping of a non-spatial dimension by the hippocampal-entorhinal circuit. *Nature* 543, 719–722
 67. Bellmund, J.L.S. *et al.* (2018) Navigating cognition: spatial codes for human thinking. *Science* 362, eaat6766
 68. Fleming, S.M. (2020) Awareness as inference in a higher-order state space. *Neurosci. Conscious.* 1, niz020
 69. Shea, N. and Frith, C.D. (2019) The global workspace needs metacognition. *Trends Cogn. Sci.* 23, 560–571
 70. Dwarakanath, A. *et al.* (2023) Bistability of prefrontal states gates access to consciousness. *Neuron* 111, 1666–1683
 71. Dijkstra, N. and Fleming, S.M. (2023) Subjective signal strength distinguishes reality from imagination. *Nat. Commun.* 14, 1627
 72. Dijkstra, N. *et al.* (2023) Distinguishing neural correlates of prediction errors on perceptual content and detection of content. *PsyArXiv*, Published online February 8, 2023. <http://dx.doi.org/10.31234/osf.io/6rs8d>
 73. Sergent, C. *et al.* (2021) Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nat. Commun.* 12, 1149
 74. Schurger, A. *et al.* (2009) Reproducibility distinguishes conscious from nonconscious neural representations. *Science* 327, 97–99
 75. Kob, L. (2023) Exploring the role of structuralist methodology in the neuroscience of consciousness: a defense and analysis. *Neurosci. Conscious.* 2023, niad011
 76. Heywood, C.A. and Kenridge, R.W. (2003) Achromatopsia, color vision, and cortex. *Neural Clin.* 21, 483–500
 77. Haynes, J.-D. and Rees, G. (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* 8, 686–691
 78. Kriegeskorte, N. and Mur, M. (2012) Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* 3, 245
 79. Roads, B.D. and Mozer, M.C. (2019) Obtaining psychological embeddings through joint kernel and metric learning. *Behav. Res. Methods* 51, 2180–2193
 80. Kleiner, J. (2024) Towards a structural turn in consciousness science. *Conscious. Cogn.* 119, 103653
 81. Brouwer, G.J. and Heeger, D.J. (2009) Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* 29, 13992–14003
 82. Brouwer, G.J. and Heeger, D.J. (2013) Categorical clustering of the neural representation of color. *J. Neurosci.* 33, 15454–15465
 83. Pollack, I. and Khouri, N. (1979) Similarity space for auditory signals differing in frequency and intensity. *Bull. Psychon. Soc.* 13, 209–211
 84. Oh, S. *et al.* (2020) Towards a perceptual distance metric for auditory stimuli. *arXiv*, Published online October 30, 2020. <http://dx.doi.org/10.48550/arXiv.2011.00088>
 85. Jraissati, Y. and Deroy, O. (2021) Categorizing smells: a localist approach. *Cogn. Sci.* 45, e12930
 86. Bremner, A.J. *et al.* (2013) 'Bouba' and 'Kiki' in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition* 126, 165–172
 87. Lyre, H. (2022) Neurophenomenal structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neurosci. Conscious.* 1, niac012