Synthese (2021) 199:5917–5935
https://doi.org/10.1007/s11229-021-03052-4

T.C. : INTERDISCIPLINARY PERSPECTIVES ON LOCATING REPRESENTATIONS IN BRAIN

Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations

Adina L. Roskies[1]

Received: 19 May 2020 / Accepted: 21 January 2021 / Published online: 9 February 2021

Abstract
Functional neuroimaging is sometimes criticized as showing only where in the brain things happen, not how they happen, and thus being unable to inform us about questions of mental and neural representation. Novel analytical methods increasingly make clear that imaging can give us access to constructs of interest to psychology. In this paper I argue that neuroimaging can give us an important, if limited, window into the large-scale structure of neural representation. I describe Representational Similarity Analysis, increasingly used in neuroimaging studies, and lay out desiderata for representations in general. In that context I discuss what RSA can and cannot tell us about neural representation. I compare RSA with fMRI to a different experimental paradigm which has been embraced as being indicative of representation in psychology, and argue that it compares favorably.

Keywords Functional MRI · Representation · Content · Realism · Naturalism

1 Introduction

In psychology and cognitive science there has been sustained debate about the metaphysics of mental representation: are mental representations objectively real, useful fictions, or nonexistent? Neuroscientists too debate whether neural machinery realizes mental representations, or whether our brains are merely dynamic causal systems that in no meaningful sense represent. Furthermore, what would count as evidence for representation in the brain? Although the groundbreaking work of Hubel and Weisel (Hubel and Wiesel 1959, 1962, 1998), whose recordings from cat and monkey visual cortex promised to reveal how neurons represent aspects of the visual

✉ Adina L. Roskies
adina.roskies@dartmouth.edu

[1] Department of Philosophy, Chair, Cognitive Science Program, Dartmouth College, Hanover, NH 03755, USA

world, and subsequent work recording from single cells in various cognitive neural systems has been taken by many to be evidence of neural representation, functional neuroimaging has been criticized as being unable to speak to such questions, telling us merely where, but not how content is represented in the brain (Coltheart 2006a, b; Fodor 1999).

I would contend that even in the early days neuroimaging was up to far more than merely locating where in the brain something happened (see Roskies 2009), but the past decade has seen remarkable advances in the sophistication of analysis and interpretation of neuroimaging data that even within the field alter our evaluation of the prospects of imaging to inform questions of cognitive science. In particular, the field has largely shifted away from identifying the primary changes in regional brain activation for cognitive tasks (univariate analyses), to looking at complex patterns of signal changes and the ways in which they relate to one another and to task demands (multivariate analyses). The change has been characterized as a change from a focus on activation to a focus on information (Kriegeskorte et al. 2006; Kriegeskorte and Bandettini 2007), and, I submit, it is a change that brings us much closer to being able to study representation in the brain.

In what follows, I will argue that fMRI data gives us a window into the structure of representations in the brain. One illustration of this comes from model-based analytical techniques. In Sect. 2 I begin with a brief primer on fMRI, and in Sect. 3 I discuss RSA in particular. Section 4 lays out the desiderata for a physical phenomenon to count as having representational content. Sections 5 and 6 explore what RSA can tell us about neural representation, and what it cannot, respectively. I conclude that RSA and neuroimaging in general provide us with proxy representational vehicles, and provide access to constructs that are provisionally representational. Is that a problem? In Sect. 7 I discuss a classic experiment in cognitive psychology, and compare it to RSA with fMRI. The last section concludes with a discussion of how RSA fits into cognitive neuroscience more broadly.

2 Functional MRI

Functional Magnetic Resonance Imaging (fMRI) is a technique that uses magnetic resonance technology to allow us to infer, roughly, the changing patterns of neural activity across large areas of cortex or the whole brain during the performance of specific tasks. Briefly, by subjecting tissue or other materials to rapidly changing magnetic gradients and radiofrequency pulses, the MR scanner can measure certain properties of aggregates of molecules in a region of tissue (Buxton 2009). fMRI measurements provide coarse-grained information regarding aggregate neural activity in cuboid regions called voxels (volumetric pixels). In most functional MR paradigms, the construct measured is the ratio of oxygenated to deoxygenated hemoglobin in blood, which correlates roughly with blood flow, which is independently known to correlate with local neural activity. Thus, through a complicated series of inferences, estimates can be made of the regional neural activity throughout the brain or in a defined segment of brain tissue (Roskies 2008). The resolution of most fMRI studies is on the order of 3 mm^3. Since millions of neurons reside in each of

these approximately pea-sized region of tissue, these measurements are unable to provide anything like complete information about the neural activity in the region – different subsets of neurons in a voxel may perform quite different functions yet give rise to the same hemodynamic response. However, systematic structure/function mappings in many brain areas allow us to make functional inferences even given the coarse level of grain of fMRI.

Building on methods developed with Positron Emission Tomography (PET), early fMRI studies compared activation levels between a task of interest and a baseline task, reporting the regions with greatest reliable differences in activation and associating these with psychological functions identified from hypothesized decompositions of task performance. Because brain tissue is always active, even when a subject is at rest, insight into the regional activity that correlates with task-specific changes in function is achieved by comparison of activity across different kinds of tasks. Initially, univariate analyses that focused on reliable observed changes in activation magnitudes aimed at localizing task-related functional changes in brain activity. Although a number of methodological improvements allowed fMRI to far surpass PET using these analysis methods, it was not until multivariate analysis methods were developed that fMRI began to show its full potential.

Multivariate analysis techniques utilize the *pattern* of responses across multiple voxels rather than focusing on localized regions of greatest change (Haxby et al. 2001). Analyses that harness the information from patterns of activation are called multivoxel pattern analyses (MVPA). For example, a set of nine contiguous voxels might have the same average activation level across three experimental conditions, so a univariate analysis would fail to identify that region as being important in comparisons across these conditions. These voxels may nonetheless show reliably different patterns of activity across conditions, which would be discernable with multivariate analyses. A multivariate analysis might show that sufficient information is encoded in each condition to discriminate it from the others. If so, then each pattern carries information about which condition the subject was experiencing when the data was generated, and, one can infer, about the neural processing specific to these conditions.

Employing machine learning classifiers on activity pattern data to classify new data on the basis of trained patterns is called MVPC[1] (multi-voxel pattern classification) (Haxby et al. 2001). Much of interest has been learned using these decoding methods. For example, in an early study, machine learning classifiers trained on multivoxel data from regions throughout occipital cortex showed that broad regions of cortex encoded sufficient information to discriminate between perception of different classes of objects. Moreover, it was unexpectedly revealed that in a region thought to be selective for face processing on the basis of univariate data, sufficient information was present in non-face conditions to allow a classifier to determine which of several stimulus types the subject was being shown (Haxby et al. 2001).

[1] MVPA, or multi-voxel pattern analysis, is an umbrella term used for all sorts of methods which involve analyzing fine-grained fMRI data. We use MVPC, or multi-voxel pattern classification, to indicate MVPA methods using classifiers (see Haxby et al. 2014).

Although MVPC has become a standard method for analyzing fMRI data, its relevance for understanding neural representation is limited. MVPC with linear classifiers reveals whether patterns of brain activity in different conditions are linearly separable. If they are, there is sufficient information in the pattern of activation at the given resolution to enable a classifier to determine which of the patterns a new piece of data is most similar to. Linear separability also suggests that this information is *explicitly* represented, in that downstream brain regions could read out this information given standard theories of population coding (Kriegeskorte and Kievit 2013), though it does not go so far as to demonstrate that they do. For example, in the face-processing study mentioned above, MVPC only indicates that the information about non-face objects is present in the FFA, it does not reveal whether that information is *used* by the brain in non-face object processing (Diedrichsen and Kriegeskorte 2017; Schalk et al. 2017). Indeed, establishing causality is a general problem for neuroimaging: Neuroimaging reveals information correlated with the stimulation conditions and behavioral outcomes of experiments, but does not prove causation. Other kinds of manipulations (i.e. interventions) are better suited to establishing causality (Schalk et al. 2017).

A second limitation of MVP classification is that successful classification is silent on important questions regarding the brain data it uses. First, it does not reveal what aspect of the data allowed for classification. Classifiers can pick up on any aspect of the training set, not necessarily the one the experimenter intends, or one that is psychologically salient for humans, or the one that the brain actually uses (Kriegeskorte and Diedrichsen 2019). In general, MVPC does not provide information about the nature of the information successful classification reflects (Ritchie et al. 2019). Second, MVPC is spatially ambiguous: Different cortical patterns in different regions can give rise to identical classification performance, and successful classification does not care or reveal where in the brain the information that allows for performance originates from (Naselaris and Kay 2015). This also limits its scope for yielding insight about neural representation. Still, spatial relevance can be interrogated in particular ways, for example, by restricting the location of training data by a searchlight procedure, which looks serially across cortex for patterns in localized regions of cortex.

3 Representational similarity analysis (RSA)

Despite its limitations, the recognition that patterns of activity within a region can contain information relevant to cognitive processing set the stage for further methodological advances. The method I will focus on for the rest of the paper is another type of MVPA, still relatively novel to neuroimaging, but with a long history in psychology. In neuroimaging it is called Representational Similarity Analysis (RSA). RSA is the calculation of pairwise similarities between patterns of response, whereas second-order RSA is a method for calculating the similarity between different representational geometries (Kriegeskorte and Kievit 2013; see below). RSA takes as input activation levels in an array of voxels, but instead of using this information for classification, it looks for internal similarity relations and their relationship to

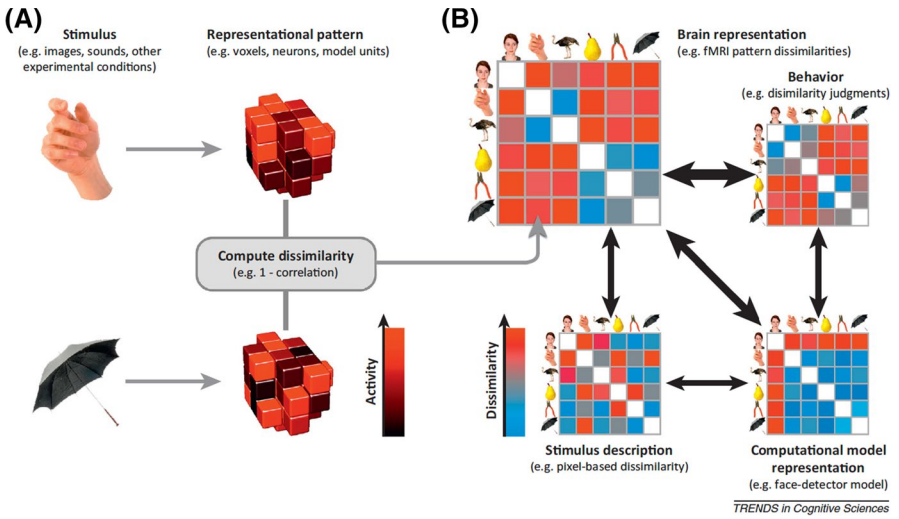

Fig. 1 (A) First order RSA: Differences between patterns of activity in a chunk of tissue responding to two objects, here a hand and an umbrella, populate once cell of an RDM in (B). (B) A complete RDM can now be compared using second order RSA with other RDMs constructed from behavior, input measures, or other models. Source: This figure is reprinted with permission from Kriegeskorte and Kievit (2013)

those of structured models that are either empirically or theoretically derived. It is thus one of a growing number of model-based fMRI analyses (Diedrichsen and Kriegeskorte 2017).

Before describing RSA in greater detail, I explain some technical terms. A representational space defines a similarity metric on measurements in a dataset, and distances in that space represent degree of similarity given a chosen metric. The set of relationships across the dataset describes a representational geometry, the geometry of that similarity space. Thus, representational geometry characterizes the relationships between a set of points in a representational space, where distance in that space is a measure of representational similarity. In a representational space for objects, one might find, for example, that all animate objects will cluster in a relatively defined region of that space, clearly separable from inanimate objects, and furthermore, that among animate objects, 4-legged creatures and winged creatures form distinct clusters (Fig. 1). In some representational spaces the ordering of objects reflects patterns that are interpretable, such as taxonomic relationships.

First order RSA allows one to represent the similarity space of one set of data in a matrix, and second-order RSA enables one to compare the similarity spaces of different sets of data in a different matrices (of same dimensionality). What is key is that, once the similarity spaces are both described in matrix form, one is able to compare representational geometries in format-neutral ways—ways that abstract away from the format of the data. Thus, one can compare for example, patterns of brain activity to stimuli or tasks across individuals, or across brain regions, or across species, or even to compare them to psychological variables such as similarity measures, reaction time,

computational models, and so on. Because RSA looks only at internal relationships, abstracting from the format of the representations themselves, and from the absolute magnitude of the measures, it allows comparisons of structure across widely different representational modes. So while classification depends upon the discriminability between patterns of brain activity, RSA characterizes the interrelations among different patterns, illuminating structural aspects of the problem.

RSA works as follows: In first order RSA, patterns of brain activity during different experimental manipulations are encoded as vectors, and a distance metric is used to calculate the distance (Euclidean, angular, etc.) between pairs of vectors. These distance measurements comprise a representational dissimilarity matrix, or RDM, of the activity in that task. The individual entries in an RDM are thus numbers characterizing the dissimilarity of the brain activity patterns across the two conditions, yielding a diagonalized matrix. As RDM entries are scalar quantities, they are content-neutral and abstract from particularities of their provenance, such as spatial layout. In RSA one then performs second-order comparisons between different RDMs. In general, one compares RDMs corresponding to different measurements over the same experimental conditions. It is this second-order comparison between RDMs that makes RSA so versatile, allowing comparisons between sets of measurements that differ in substantial ways. In sum, RSA provides a means of comparing structural relationships embedded in multivariate data to the structural relationships in other datasets.

To give a brief example, one could construct an RDM from fMRI data of inferotemporal cortex (IT) activity patterns from a human observer looking at pictures of different animals, and another RDM from the spike train data from neurons in IT in a monkey looking at the same array of animal pictures. Each of those RDMs show certain kinds of similarity structures (Fig. 2), namely similarities that closely respect taxonomic categories (Kiani et al. 2007; Kriegeskorte et al. 2008a, b). Comparison of the two RDMs shows that the two organisms have very similar representational geometries for animal pictures in IT, despite the fact that the data is generated in different species with different methods at very different spatial scales (Kriegeskorte et al. 2008a, b).

Importantly, the term ‘representational’ used in ‘representational space’ and ‘representational geometry’ refers to the technical mathematical usage of the term representation, rather than the term used in cognitive psychology or philosophy of mind. Nonetheless, the word in the context of neuroimaging plays a dual role, because we assume that the neural states that are causally involved in the performance of cognitive tasks realize the kinds of mental representations that are involved in task performance described at a psychological level. The importance of similarity metrics and representational geometries for studying representation has been recognized for decades in psychology and related fields. But we cannot take for granted that points in a representational space in the mathematical sense are themselves representations in the cognitive sense. The question at issue is to what degree are the objects in the representational space defined by fMRI activation representations in the cognitive sense?

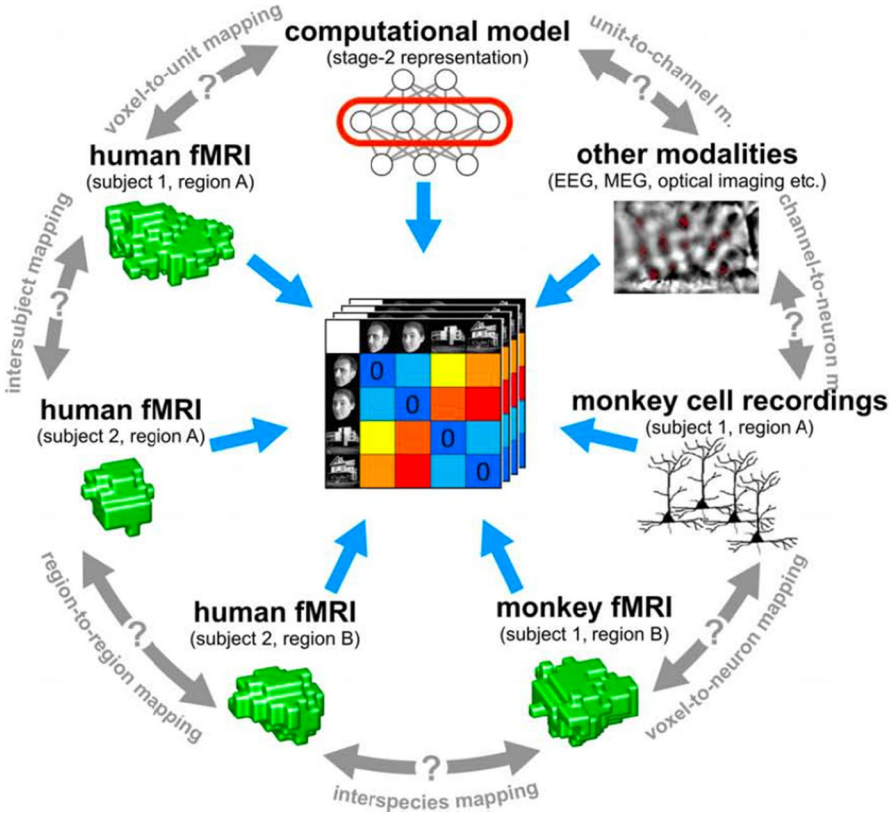

Fig. 2 Comparing RDMs through second-order RSA can relate representational geometries across very different methods, whereas direct relationships can be difficult to establish. Source: This figure originally appeared in Kriegeskorte et al. (2008a)

4 What is representation?

Neuroscientists and philosophers have been concerned with the nature of mental representation and the way(s) in which mental representations could be instantiated in brain tissue. The most promising framework for understanding cognitive function is computationalism, still the driving insight into thinking about how a physical system can perform complex tasks. Computationalism takes cognition to be a kind of computation, where computations are operations over representations.

A computational explanation of a given neural function would require an account of the representations involved in the computation (an elucidation of their structure and their content—the interpretation function), as well as an account of how these representations are physically realized, and the causal processes that operate on these representations to generate the appropriate output (the realization function) (Egan 2010, 2018).

Any computational explanation involves several substantive commitments. First, it is committed to a distinction between representational vehicle and representational content. Representational vehicles are physical structures or states that carry representational content (Egan 2018). In a computational system, causal processes act upon the representational vehicles to effect physical transformations, and in so doing the representational content of the representations is also altered. Parallels between the formal or physical properties of representations and their semantics make possible intelligent behavior.

In addition to a commitment to representational vehicles and content, we can articulate several other desiderata for candidates for representation in the brain (see, e.g. Egan 2020; Ramsey 2007). First, **function**: representations must have a function in the system. In Clark's minimal representationalism, he calls "a processing story representationalist if it depicts whole systems of identifiable inner states (local or distributed) or processes (temporal sequences of such states) as having the function of bearing specific types of information about external or bodily states of affairs." (Clark 1996). Thus, representations must be part of a system of representations with a function for the organism. A corollary of this is that representations must be *used, consumed* or *interpreted* by the system, and they must play a role in explaining intelligent behavior. In addition, content ascriptions must be **naturalizable**: both the realization and interpretation functions must be able to be cashed out in physical matter and processes. Finally, as many have noted, representation implies the possibility of **misrepresentation**. At the very least, there must be norms for content ascription, even if these norms are relative to the workings of the system. Norms, however unnaturalizable they may seem at first glance, are implicit in attributions of function (see first desideratum), and violated when the functional goals are not achieved.[2]

4.1 From mental content to neural representation: a proposal

All philosophical theories of mental content rely on the idea that the brain must allow the organism to track and respond in adaptive ways to the environment. The main accounts philosophers have offered for this tracking relation are causal or information-theoretic theories, which anchor the content of a representation in a causal story; teleological theories, which ground content in a story of natural or evolved function; and structural theories, in which isomorphism or similarity plays a role in relating referent to representation. Although each approach has its shortcomings or problem cases, several philosophers have recently proposed that a unitary account of content assignment is unnecessary, while other theorists have stressed the

[2] Additional desiderata have been suggested for vehicles, at least for the vehicles of representations in a digital system: Clark notes that vehicles need to be "portable", in that the same vehicle can play a role in different computations, and that they be "type-able", or able to be classified into types. These desiderata are more contestable, and seem to be suited to specifically digital representations, as properties of representations that operate in a digital system. It seems to remain a possibility that analog representations, like pictures, be neither typeable or portable in the way they must be in a digital system, yet that they are legitimate representations, and perhaps even paradigmatic representations.

pragmatic or explanatory role that content ascriptions play in our scientific practice (Egan 2020; Godfrey-Smith 2006; Shea 2018; Shea et al. 2018). I am more sympathetic to a realist interpretation than some, for, as Shea says, "What adverting to content does achieve, however, is to show how the system connects with its environment: with the real-world objects and properties with which it is interacting, and with the problem space in which it is embedded." (Shea 2013, p. 498). Indeed, it is this feature that makes it the case that representations play functional roles. One can, however, take a realist view yet allow that content identification and explanation can appeal to different theories in different cases.

Despite the unsettled nature of theories of content, there is little doubt that there are causal processes that relate brain activity to external stimuli, and that play a role in transformation of early perceptual representations to more articulated representations higher in the processing stream. Recent decades have seen significant progress in theorizing about how to characterize such representations. One promising approach involves thinking of brain states as high-dimensional state-spaces. Briefly, the brain's immense complexity endows it with the capacity to represent arbitrary combinations of elements in a high-dimensional space. Some of these dimensions are straightforwardly represented in the spatial topography of the cortex (for example, the somatotopic representations in sensory and motor cortex, and the visuotopic layout of early visual cortices), while others, such as orientation-selectivity, are iterated in substructures like cortical columns, and yet others seem to be encoded purely functionally. Evidence from single unit recording suggests that representational properties are organized hierarchically within cortex, with neurons in lower sensory and motor cortices representing simpler properties and more complex semantic properties emerging in higher cortical regions via iterated transformation.

One of the reasons to frame neural coding in terms of high dimensional spaces is that this framework easily adapts to both content and neuroimaging domains (Haxby et al. 2014). Consider a representational space in which each dimension represents specific feature of a given problem area. Let us call this a *semantic space*. Consider, for example, the set of features that characterize objects. Relevant dimensions might be size, colors, shape features, and so on. Each object will be represented as a point in that space. Of course, we are interested in neural representation, and in particular how neuronal processing can underlie cognition and behavior. We can conceive of a *neural representational space* as a very high multidimensional space, with (for example) each neuron in a region represented by a different axis in that space, and the activity of all neurons in that region at a time defining a particular point in that space. Supposing that region is a region involved in visual processing, responses of that population of neurons to different visual stimuli will correspond to different points in that space, and we could examine how that activity vector moves around in the neural representational space given defined changes in the visual stimulus. To link this to the basic notion of computation, understanding neural representation (say, in a brain region) would involve understanding the dimensions of the semantic space there represented (equivalent to the interpretation function), and the mapping

from the neuronal representational space to the semantic (equivalent to the realization function).

Although fMRI gives us a measure of brain activity, it does not give us access to the individual activity of neurons in a brain region, but rather to roughly the aggregate activity of neurons in a defined region of brain tissue. We could therefore construct another multidimensional space in which the activity of each voxel mapped to an axis in that space. Of course, the dimensionality of such a space would be much smaller than the neural representational space, and many different axes of the former space would be collapsed into the latter. If all neurons in a voxel were tuned to different individual features and the set of stimuli ranged over all these features, one would not expect to see differences across different stimulus conditions. However, we know that neurons with similar functions tend to be spatially located near one another, and we do tend to see reliable differences across different conditions in many brain areas, indicating some selective tuning to task-related parameters, even at the relatively coarse resolution of fMRI. For example, Kamitani and Tong (Kamitani and Tong 2005; Tong et al. 2012) were able to determine the orientation of a stimulus from patterned fMRI activity, despite the fact that voxels are orders of magnitude bigger than the cortical columns that encode orientation. Their success was based upon net signal arising from local anisotropies in the cortex. Just what and how much task-related information can be discerned at the level of the voxel is unknown, and it is sure to differ between areas, and to depend upon scan resolution.

5 What does RSA tell us about neural representation?

Does RSA identify representations in the brain? The answer here is: partially. We can begin by asking whether there is a distinction between vehicle and content. A simple reading would say yes: the voxelwise activity pattern is the vehicle, and the content is illuminated both by looking at the external cause (or effect) of the neural activity as well as by examining the covariation of the pattern with the content of the structure to which the activity is being compared (i.e. the target of the RSA). In this sense the primary metric of the second-order comparison is iso/homomorphism, for the degree to which the content is similar in structure to the domain with which it is being compared is indicative of its representational content, derivative on the content ascribed to that structure. The content ascribed to the primary representations is often ascribed on the basis of a theory of content, such as one (or several) described above. As in cognitive science more generally, in neuroimaging the theory by which content is ascribed is assumed rather than explicated.

For a realist about mental representation, however, voxel activation values are not a proper vehicle of content, for they do not meet the realist criteria necessary to support a computational role. Representational vehicles must have causal powers, and voxel values emphatically do not. The real vehicles of content must be select

neural subpopulations that are responsible for the anisotropies of activation seen across conditions. Thus, in neuroimaging we do not get direct access to vehicles of content, but must infer their presence, from the observation of reliable, theoretically-meaningful (seeming) patterns of activation. Voxelwise activities are only proxies for representational vehicles: they tell us where to look for them with other methods, but do not individuate them.

Even if we only get vehicle proxies, RSA does prove informative about content insofar as content is encoded or reflected in structural relationships.[3,4]RSA with neuroimaging suggests how representational machinery is structured across the cortical surface at a macroscopic level, and is able to discern multiple representational spaces within the same tissue, in line with hyperdimensional theories of representation. In some instances, these suppositions get confirmation, as in when the results delivered with single-unit recording corroborate RSA results. For example, in a study of monkey and human IT, the similarity matrix among a large stimulus set of animals was strongly isomorphic to that derived from individual neuronal firing rates in monkey IT when monkeys were exposed to the same stimulus set (Kriegeskorte et al. 2008a, b). The implication is that the large-scale organization of neurons representing the animal hierarchy reflects the finer-grained cellular receptive field properties. The extent of these similarities suggests that both species represent categorical structure among objects similarly. And as we have independent reason to think that our early visual systems are highly homologous in terms of basic neural function, the fact that the structural relationships are so similar in IT provide further reason to infer that vehicles for object representation in this area is similar as well. Among other things, this provides (defeasible) reason to interpret single cell level neural data gathered from monkeys, which we can obtain, as informative about neural coding in human IT, information which in general, we cannot obtain directly.

One might wonder why blood flow can be a proxy vehicle for what is in reality a much finer grained vehicle (most likely firing rates or patterns of select populations of neurons, whose precise characterization we do not yet know). It may also seem puzzling that there are similarity relationships between brain measurements as different as those of individual cell responses to stimuli and of aggregate activity in that region to the same stimuli. It certainly could be otherwise: one could imagine that even if neural firing rates reflect categorical similarity, the way these neurons are distributed within regions of cortex might be relatively homogeneous, so that at the resolution of fMRI voxels, no relationships are found between blood flow changes and categorical structure. However, these structural parallels apparent between different forms of measurement would be less surprising if the brain uses topography in a multidimensional space as a way of organizing and processing content (so, for

[3] RSA yields results which are informative without being determinative: We can assess representational structure relative to a hypothetical model or relative to competing hypotheses, but cannot rule out other interpretations that share representational geometries.

[4] Although the vehicle and content questions are logically separable, in practice they are interdependent. You cannot investigate content without identifying the vehicles that carry the content, but a way of identifying the vehicles in natural systems is to look for structures embodying content-relevant relationships in causal pathways that are candidates for representing.

example, that similar content undergoes similar processing), and if projection of that multidimensional space onto the voxel space is nonhomogeneous. If these conditions hold fMRI signals can reflect content-related structure, and thus also function as a proxy vehicle for representations. In other words, widespread structural similarity across brain regions and across species suggests that the brain harnesses representational structure in its computational solutions to the problems of cognition at multiple scales.

Another way RSA could be leveraged to inform us about processing is to suggest what kind of transformations occur between closely related stages of processing. If RSA reveals a stimulus or task feature that at one stage of processing contributes to differences in the similarity space, but at a next stage that feature appears as an invariant, we can make inferences about the underlying computations and/or intervening representations. For example, in early visual cortex face stimuli do not cluster together in RDMs, but in higher levels of the visual pathway, such as IT, they form a distinct similarity cluster. In a later stage, similarity measures for individual faces do not differ even when the face stimuli are presented in different orientations (Guntupalli et al. 2017), suggesting that identity is computed between these stages of the visual hierarchy.[5] By allowing us to probe which kinds of stimuli or behaviors result in invariances in the similarity matrix, and to look for the emergence of such invariants, we can infer where and when in the processing hierarchy certain higher-order properties are computed/extracted from the signal. When these invariants are part of a system that subserves behavior and are correlated with observed behavioral capacities, we have a prima facie case for **function**.

Although the content of neural representations can sometimes be understood by reference to objects causally related in the world, such as in the case of object recognition in IT cortex, sometimes the content is harder to specify – neural content need not be semantically transparent. In those cases content is better ascribed by its functional role. Searching for identifiable transformations or the emergence of invariants are ways of doing this. Although low level features or intermediate representations may not map neatly onto the way in which we categorize or conceptualize things, similarities in RSA, and also in deep learning networks and intermediate-level brain regions, have representational status because they are part of system that mirrors or encodes structural features of the world, reducing the effective search space, even if their content is not easily parsed or identified (see Poldrack, this issue).

The problem of **misrepresentation** is more difficult, but tractable nonetheless. In the philosophical literature it has been claimed that no sense can be made of misrepresentation without the ability to assign determinate content. As noted, functional accounts and structured maps imply norms: the system must be systematically organized to function properly. The fact that representational geometries tend to be found across persons and even species, and that they can be elicited in many different brain areas gives us a handle on understanding misrepresentation. Finding

[5] Here we have evidence from anatomy that there are stages of processing hierarchically organized, and thus can infer that representations at one stage causally affect the next. But fMRI does not provide direct evidence of causal connection.

discernable representational geometries across regions, combined with knowledge of microanatomical local structure gives us reason to posit a norm. If, for example, there is a gradient of predacity in IT (Connolly et al. 2012), and a stimulus of a mouse leads to activation in the high-predacity area, along with lions, tigers, and bears, yet it was situated properly in the representational geometry at other levels of the hierarchy, we could say the subject misrepresented the mouse as being dangerous.[6] This interpretation would be strengthened if we found the representation played a role in behaviors at odds with the agent's goals (Isaac 2013).

RSA provides a means of characterizing content, and thus postulating determinate content, especially in the second-order comparison, which can be chosen by the experimenter. Systematic deviations could indicate a lack of correspondence to these content ascriptions, whereas punctate deviations in individuals could indicate misrepresentation.

To sum up, RSA indicates that we can infer representational contents on the basis of proxy vehicles. It further suggests that the cortex encodes some semantic features in a map-like way, with semantic relationships mirrored in similarity of patterned activity. Moreover, computational transformations can be inferred by comparing contents across regions. Interestingly, this suggests that, at least at the scale of fMRI, neural representation looks more analog than digital, contra the kind of computational picture espoused by classic computationalists, such as Fodor. Rather than contents being arbitrarily associated with symbols, contents are structured and represented systematically in cortical regions, allowing inferences about representation on the basis of representational geometry.

6 Provisional representations: What doesn't RSA tell us

I call the kinds of representational structure indicated by analysis of fMRI activation patterns *provisional representations*. They have many of the qualities of mental representations as realized in neural hardware, yet on some crucial dimensions there is just a promissory note. Importantly, RSA yields information about representations that leaves open the nature of the vehicles of content. We know the vehicles of content are not voxels, since these are mere constructions of the experimenter. We do not know which populations of neurons are responsible for carrying the content elucidated in the similarity matrices, and what other information they carry that is inaccessible to us at the spatial scales that fMRI provides. How much of a problem is that? Some might argue that unless we have access to the actual vehicles, we cannot make a case for representations: as Egan has said, "No vehicles, no representations." She has allowed that vehicles can be inferred,[7] and thus one can still make the case for representations with presumptive or proxy vehicles. RSA tells us that the brain regions contributing to the analysis carry the relevant information at the scale of fMRI.

A second, perhaps more troubling issue, is that we don't know whether or how the information that informs the RSA is read out by later processes. Although RSA

[6] This presupposes that there is a readout that respects the representational geometry.

[7] Egan, personal communication.

demonstrates that structural information is available in a signal, it does not guarantee that that information is made use of by downstream processes. It is this that seems to constitute the biggest failure of RSA to provide access to representations, for we lack even a candidate for a proxy in this respect. It is an existence proof for the presence of the relevant information, but not of its use. Unless information is *consumed* by the downstream processes, it is epiphenomenal to behavior. And even if we have evidence that information of this sort is available to later processes, we do not know that the read-out mechanisms operate on the scale picked out by the resolution of the fMRI study.

The reservations we have due to this limitation can be assuaged by tying our experiments closely to behavior. For example, Charest et al. (2014) performed RSA on fMRI data from IT when subjects were shown familiar and unfamiliar objects. Similarities between the representational geometries were clear for all subjects. Moreover, they showed that individual differences in the representational geometries predicted individual differences in similarity judgments by those individuals for familiar objects. They hypothesized that the representation of these objects was due to idiosyncratic experiences of the subjects with the objects, which also underlay their similarity judgments. While causality is not demonstrated here, it is inferred from the close correlation with individual differences. Similar causal inferences could be made by treating an experimental manipulation as the intervention, and seeing if and where similarity emerged in RSA.

The availability of correlational evidence without evidence for causal efficacy is one that plagues fMRI and other neuroimaging studies more generally: in the absence of interventions, we lack proof that the activations we see reflect information that is causally relevant. That is, fMRI investigations of representation give us a "how possibly, but not necessarily a "how actually" story. Although I fully acknowledge this limitation, I suggest we recognize that no tool in the neuroscientist's toolkit is without limitations, and this work presents us with a clear hypothesis to test with other techniques. Evidence of causal involvement is thus a crucial part of the larger scientific project.

Indeed, it is worth pointing out here that the multidimensional framework outlined here is useful for analysis of convolutional neural nets (CNNs) or deep learning networks, and that structures in such networks trained to do cognitive tasks that humans and animals do are similar to those inferred from neural data (Khaligh-Razavi et al. 2017). Thus, insofar as these structures in CNNs are taken to be evidence of representation, so should representational geometries found in fMRI data (Poldrack, this issue; Yamins et al. 2014; Yildirim et al. 2019). One difference is that causality is relatively easy to evaluate in CNNs, in contrast to fMRI. Thus, evidence for causality of representations in CNN models may be useful in arguments for causality with regard to provisional representations in fMRI data.

7 Representations in psychology

As we have seen, RSA gives us some reason to attribute representational status to fMRI measurements, but fails to satisfy all desiderata. In case one is tempted to deny the patterns seen in fMRI representational status because of these lacunae, it would

be useful to compare the fMRI results with other data in psychology that seems to unquestionably embrace representational talk. Consider, for example, any of a number of commonly used methods in cognitive psychology that measure reaction times or some other behavior (eye movements, for example) in order to make inferences about the existence of mental representations and/or their properties. These kinds of experiments, it seems, while central to the discipline, are on no firmer footing than RSA is in revealing the inner workings of the mental. Take, for example, the work on mental rotation by Shepard and Metzler (Shepard and Metzler 1971) and subsequent studies using similar paradigms.[8] In their initial studies, Shepard and Metzler found that the time it took for subjects to judge whether two two-dimensional visual projections of three-dimensional geometric objects were of identical or mirror reversed objects, scaled linearly with the angle of rotation, and, perhaps more interestingly, reaction times were similar regardless of whether the requisite rotation was within the visual plane or involved the third (depth) dimension. These reaction-time measurements led them to posit the imaginative construction and mental rotation of 3-dimensional mental shape representations. Their initial results were taken to be of sufficient interest that the study made the cover of *Science*. Although the initial paper did not mention representation, but rather phenomena that could be explained by reference to representation such as imaginings of three dimensional objects, later interpretation took these results as evidence for analog and spatial (image-like) mental representations, and what Shepard termed a "second-order isomorphism" between image and object. While these have not been immune to criticism (Carpenter and Just 1978; Just and Carpenter 1976; Pylyshyn 1973), the criticisms have generally been of the form that ordinary scientific discourse has taken, concerning the characteristics of these mental objects, and ultimately leading to new experiments to rule out alternative interpretations (Shepard and Cooper 1982) rather than to dismissal of their scientific relevance, or to discussion about whether the data supports the existence of such mental representations. Indeed, subsequent experiments also relied upon reaction time methods.

Notice, however, that the Shepard data gives one no access to representational vehicle either. One merely infers that there is a 3D mental construct that the subject rotates, though one can calculate from the data the limits to the speed of rotation. In this way, the chronometric analysis is even more removed from the representation itself than is data from RSA with fMRI. Reaction times give one no access to the vehicle of representation at all. Instead, they give one a reason to posit a certain kind of content (a 3D as opposed to 2D shape) and perhaps something about the nature of the representational vehicle (image-like as opposed to digital).[9] The positing of a representation is abductive.

[8] Interestingly, it was Shepard who was one of the first to apply representational geometry methods in psychology, using similarity as a way to characterize content (see Shepard 1987).

[9] The Shepard data has often been interpreted as indicating that the relevant mental representations are image-like. But in the multidimensional framework discussed here, the RT data could be interpreted as being consistent with the dynamics of a movement of an activation vector through a space homologous with 3-D rotational space. In such a framework the clear distinction between image-like and digital or discursive tends to fall apart.

Similarly, RSA of fMRI data provides correlational evidence for the existence of a structure with structural similarities to behavioral measures. But in addition to what the Shepard method provided, RSA points to a brain area as a plausible candidate for the vehicle—an embodiment of what in Shepard was merely an existence proof.

What differs between the mental representations inferred by Shepard and those posited with RSA is that the posited representations in the Shepard case are causally implicated in the explanation of behavior. Reaction time on a trial-by-trial basis is argued to be dependent upon the angular distance between the two stimuli: the mental manipulation of the mental construct that causes the RT curve to vary as it does. The same can be said for the Charest et al. study mentioned above, in which individual differences in similarity judgment correlated with individual differences in representational geometry. In contrast, most uses of RSA do not necessarily compare a neural representation directly to a behavioral measure. For example, the comparison between the representational geometries of the two face patches discussed earlier do not involve specific behaviors, though they make reference to a behavioral capacity demonstrated by the organism: the ability to recognize individual faces regardless of their angle of presentation. And sometimes RSA is explicitly used to compare neural representations to trial-by-trial behavioral variations. For example, when patterns of activity in IT during object recognition are found to correlate closely with explicit similarity judgments, the implication is that the behavioral measures reflect computations reliant on those representations. Perhaps what that points to is a suggestion that more effort be made to link representational geometries to one another and to behavioral measures. But even in the absence of such evidence, it seems such similarity is defeasibly evidence of representation.

8 Conclusion

Egan (2018) discusses the important role in cognitive science for using representation ascriptions as a tool for discovery. This is perhaps the primary role for RSA in cognitive neuroscience. As an example, a number of regions that respond selectively to faces have been found in human visual cortex. All have been hypothesized to play a role in face processing, and some, such as the FFA are causally implicated in face processing as well, with reports of prosopagnosia resulting to lesions in this area, and stimulation leading to changes in phenomenology of face perception, but not object perception (Parvizi et al. 2012; Schalk et al. 2017). RSA has indicated, too, that view invariance is computed at a specific level of the cortical hierarchy. It will fall to researchers with more fine-grained techniques at their command to elucidate the neural codes at both these levels, and, we hope, the underlying computations that transform face-related information between these levels. RSA has also been used to identify homologous areas in monkey cortex, which will allow such targeted explorations. In other words, RSA is exceptionally good at identifying candidate hypotheses and candidate regions for more low-level, concrete neuroscientific work on representational properties. But the identification of vehicles and specific contents will require substantiation by other methods.

It is a mistake to measure the value of neuroimaging purely in what can be concluded on the basis of neuroimaging studies alone. Neuroimaging is one tool in a growing toolkit of diverse and ever-more-powerful techniques, none of which yields all the information one would need to understand the neural basis of a cognitive process. What neuroimaging can provide is a type of behavior-related information that is very difficult to get with any other available methods. Despite its centrality, it is not information that stands alone, both in that it cannot be interpreted without relying on information from other techniques and modalities, and in that while it can direct future research using other methods, it cannot supplant it.

Thus, although RSA does not provide all the information we would need to identify the neural basis of psychologically-potent mental representations, it gives us candidates for these representations, and allows for targeted hypothesis-driven neuroscience with more fine-grained techniques. If one thinks that Shepard's famous mental rotation experiments are psychologically interesting, one should also think model-based fMRI analyses to be.

Acknowledgements This paper has benefitted from feedback from the fellows at the University of Pittsburgh Center for Philosophy of Science, from Jim Haxby, and from the participants of Neural Mechanisms Online, especially Charles Rathkopf, Dan Weiskopf, Matteo Grasso, and Michael Anderson. The work was supported in part by a fellowship from University of Pittsburgh Center for Philosophy of Science.

References

Buxton, R. B. (2009). *Introduction to functional magnetic resonance imaging: principles and techniques* (2nd ed.). Cambridge: Cambridge University Press.

Carpenter, P. A., & Just, M. A. (1978). Eye fixations during mental rotation. In J. W. Senders, D. F. Fisher, & R. A. Monty (Eds.), *Eye Movements and the higher psychological functions* (pp. 115–133). New Jersey: Erlbaum.

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.

Clark, A. (1996). *Being There*. Cambridge: MIT Press.

Coltheart, M. (2006a). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323–331.

Coltheart, M. (2006b). Perhaps functional neuroimaging has not told us anything about the mind (so far). *Cortex*, *42*, 422–427.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., et al. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience, 32*(8), 2608–2618. https://doi.org/10.1523/JNEUROSCI.5547-11.2012.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology, 13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508.

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science Part A, 41*(3), 253–259. https://doi.org/10.1016/j.shpsa.2010.07.009.

Egan, F. (2018). The nature and function of content in computational models. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind*. Abingdon: Routledge.

Egan, F. (2020). A Deflationary Account of Mental Representation. In J. Smortchkova, K. Dołęga, & T. chlicht, T. (Eds.), *What are Mental Representations?* Oxford University Press.

Fodor, J. (1999). "Diary" London Review of Books, September 30. https://www.lrb.co.uk/the-paper/v21/n19/jerry-fodor/diary.

Godfrey-Smith, P. (2006). Mental representation, naturalism, and teleosemantics. In *Teleosemantics: new philosophical essays* (pp. 42–68). Oxford: Clarendon Press.

Guntupalli, J. S., Wheeler, K. G., & Gobbini, M. I. (2017). Disentangling the representation of identity from head view along the human face processing pathway. *Cerebral Cortex (New York, N.Y.: 1991), 27*(1), 46–53. https://doi.org/10.1093/cercor/bhw344.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148,* 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837.

Hubel, D. H., & Wiesel, T. N. (1998). Early exploration of the visual cortex. *Neuron, 20*(3), 401–412. https://doi.org/10.1016/S0896-6273(00)80984-8.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8,* 441–480.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8*(5), 679–685. https://doi.org/10.1038/nn1444.

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology, 76*(Pt B), 184–197. https://doi.org/10.1016/j.jmp.2016.10.007.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296–4309. https://doi.org/10.1152/jn.00024.2007.

Kriegeskorte, N., & Bandettini, P. (2007). Combining the tools: Activation- and information-based fMRI analysis. *NeuroImage, 38*(4), 666–668. https://doi.org/10.1016/j.neuroimage.2007.06.030.

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., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, 103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103.

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.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,*. https://doi.org/10.3389/neuro.06.004.2008.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences, 19*(10), 551–554. https://doi.org/10.1016/j.tics.2015.07.005.

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S., & Grill-Spector, K. (2012). Electrical stimulation of human fusiform face-selective regions distorts face perception. *Journal of Neuroscience, 32*(43), 14915–14920. https://doi.org/10.1523/JNEUROSCI.2609-12.2012.

Pylyshyn, Z. W. (1973). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin, 80,* 1–25.

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

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science, 70*(2), 581–607. https://doi.org/10.1093/bjps/axx023.

Roskies, A. L. (2008). Neuroimaging and inferential distance. *Neuroethics, 1*(1), 19–30. https://doi.org/10.1007/s12152-007-9003-3.

Roskies, A. L. (2009). Brain-mind and structure-function relationships: A methodological response to Coltheart. *Philosophy of Science, 76*(5), 927–939.

Schalk, G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., et al. (2017). Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proceedings of the National Academy of Sciences, 114*(46), 12285–12290. https://doi.org/10.1073/pnas.1713447114.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass, 8*(5), 496–509. https://doi.org/10.1111/phc3.12033.

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

Shea, N., Godfrey-Smith, P., & Cao, R. (2018). Content in simple signalling systems. *The British Journal for the Philosophy of Science, 69*(4), 1009–1035. https://doi.org/10.1093/bjps/axw036.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge: MIT Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701.

Tong, F., Harrison, S. A., Dewey, J. A., & Kamitani, Y. (2012). Relationship between BOLD amplitude and pattern classification of orientation-selective activity in the human visual cortex. *NeuroImage, 63*(3), 1212–1222. https://doi.org/10.1016/j.neuroimage.2012.08.005.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America, 111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yildirim, I., Wu, J., Kanwisher, N., & Tenenbaum, J. (2019). An integrative computational architecture for object-driven cortex. *Current Opinion in Neurobiology, 55,* 73–81. https://doi.org/10.1016/j.conb.2019.01.010.