Deep convolutional neural networks are not mechanistic explanations of object recognition

Bojana Grujičić
bojana.grujicic@maxplanckschools.de

Max Planck School of Cognition, Leipzig, Germany
Humboldt-Universität zu Berlin, Berlin School of Mind and Brain, Berlin, Germany
University College London, Department of Science and Technology Studies, London, UK

Abstract

Given the extent of using deep convolutional neural networks to model the mechanism of object recognition, it becomes important to analyse the evidence of their similarity and the explanatory potential of these models. I focus on one frequent method of their comparison – representational similarity analysis, and I argue, first, that it underdetermines these models as how-actually mechanistic explanations. This happens because different similarity measures in this framework pick out different mechanisms across DCNNs and the brain in order to correspond them, and there is no arbitration between them in terms of relevance for object recognition. Second, the reason similarity measures are underdetermining to a large degree stems from the highly idealised nature of these models, which undermines their status as how-possibly mechanistic explanatory models of object recognition as well. Thus, building models with more theoretical consideration and choosing relevant similarity measures may bring us closer to the goal of reaching a mechanistic explanation.

Key words: deep neural networks, explanation, mechanisms, representation, object recognition, similarity measures

1 Introduction

Although deep neural networks had their breakthrough in the engineering field of computer vision (Krizhevsky et al., 2012), there has been a lot of research in recent years employing deep neural networks directly in the pursuit of neuroscientific goals. This new framework at the intersection of deep learning and neuroscience aims to offer a novel methodology for neuroscience in contrast to the traditional one (Nastase et al., 2020; Richards et al., 2019), having hopes of fulfilling not just its predictive but its explanatory goals as well (Lindsay, 2021; Cichy & Kaiser, 2019; Kietzmann et al., 2019; Kriegeskorte, 2015).

In the domain of visual neuroscience, there has been an array of findings suggesting that inner activations of deep convolutional neural networks (DCNNs) trained for an object recognition task enable predicting neural response properties in the ventral stream to a certain extent (Lindsay, 2021; Bashivan et al., 2019; Cichy et al., 2016; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). When it comes to their accuracy on the object recognition task, DCNNs are on the human performance level. Based on these findings DCNNs are said to be the most predictively successful models of the ventral stream in human and nonhuman primate brains for object recognition (Cao & Yamins, 2021a, 2021b; Storrs et al., 2021; Yamins et al., 2014). Importantly, this should not be taken as a claim that the learning mechanisms of DCNNs and the brain are similar. The claim is limited to the correspondence of the brain and DCNN processing in object recognition once DCNNs are trained.

Given these findings, the question arises whether DCNNs also provide an explanation of our capacity for object recognition. According to some neuroscientists, DCNNs are somewhat explanatory in virtue of capturing some behavioural data and explaining neural variance (Lindsay, 2021; Kietzmann et al., 2019; Kriegeskorte, 2015). In addition, philosophical interest in DCNNs has been rising lately (Kieval, 2022; Cao & Yamins, 2021a, 2021b; Buckner, 2019, 2018), with several arguments offered for the claim of DCNNs being mechanistic explanations of object recognition. Cao & Yamins (2021a, 2021b) argue that DCNNs satisfy the model-to-mechanism mapping requirement put forth for mechanistic explanations (Kaplan & Craver, 2011) – DCNNs are already sufficiently similar to their neural targets. Buckner (2018) suggests a mechanism both DCNNs and the visual cortex instantiate, expanding on Stinson's (2018) analysis of connectionist models as explanatory in virtue of capturing some generic properties of mechanisms responsible for target cognitive phenomena.

One of the most prominently used frameworks for comparing DCNN and brain activations is representational similarity analysis (RSA), forming an important evidential basis for the claim that DCNNs are mechanistic explanations. RSA quantifies second-order similarities between dissimilarities of stimuli-elicited patterns in DCNNs and the ventral stream (Kriegeskorte et al., 2008a). First, I argue that RSA underdetermines DCNNs as how-actually mechanistic explanations since a variety of similarity measures is used within the framework. Focusing on correlation and Euclidean distance, I show that they pick out different types of mechanisms in DCNNs and the brain in order to compare them, and I argue that there is a problem of relevance of these measures for the explanandum capacity of object recognition. Since there is no arbitration between similarity measures in terms of relevance, RSA underdetermines DCNNs as how-actually mechanistic explanations. Second, the highly idealised nature of current DCNN undermines their status as how-possibly mechanistic explanations of object recognition – because models are made with little theoretical consideration, the kinds of solutions they arrive at are under-constraining for the purposes of learning about the ven-

tral stream mechanism. This makes the underdetermination due to the application of different similarity measures rampant.

After describing the explanandum capacity of object recognition and saying something about DCNNs in general in the next section, I outline how mechanistic abstraction is applicable to DCNNs in section 3. I then introduce RSA as a method of comparing DCNNs and the brain in section 4 and argue in section 5 that different similarity measures pick out different types of mechanisms on the level of representational vehicles. Section 6 presents a limitation on the use of RSA to correspond representational content across systems. I then go on to argue that DCNNs are neither how-actually mechanistic explanations in section 7, nor how-possibly mechanistic explanations in section 8. Section 9 concludes.

2 Explanandum capacity of object recognition and the DCNN-based explanans

2.1 The explanandum – object recognition

There are various facets of object recognition as one of the critical capacities enabling successful interaction with the world around us (Bracci & Op de Beeck, 2022). In what follows I focus on a constrained version of what this capacity amounts to, called core object recognition that, while not being theoretically unobjectionable (cf. Bracci & Op de Beeck (2022)) is the notion often used at the intersection of deep learning and neuroscience. I follow DiCarlo et al. (2012) in characterising core object recognition as the capacity to assign labels to objects, e.g. "orange" to an orange. Labels can range from precise ones, in which case the task is that of identifying objects, to coarse-grained ones when the task is to categorise them. Our ability to identify or categorise objects persists over various contingent conditions of presentation of the object – we are able to say an object is an orange in different lighting conditions, from different perspectives, being closer or further away from it, etc. Thus, being able to recognise objects demands solving the problem of invariance to these idiosyncratic aspects of the presentation of an object (Kreiman, 2021; Pinto et al., 2008).

Additionally, we are able to distinguish stimuli whose retinal activations can be quite similar, such as a lemon and a tennis ball. Doing this demands being able to track object-specific properties. These two requirements – invariance and specificity form the crux of the problem of object recognition (Riesenhuber & Poggio, 2000). The mechanism that can solve it needs to have a way of responding differently to retinally similar stimuli while reliably recognising these objects under different conditions of their presentation.

The task that is used to assess the core object recognition abilities consists in being presented with two-dimensional images of objects with a request to output a label for it. This task was adopted into neuroscience from the computer vision field and the ImageNet classification challenge (Deng et al., 2009). While it is debatable how representative of our object recognition abilities this task is, it is a task humans can do, and current DCNNs can solve it on the level of human task performance (Storrs et al., 2021).

2.2 The DCNN-based explanans of object recognition

Deep neural networks are computational models that consist of multiple layers of nodes connected by patterns of weights which determine the strength of the activity propagated from one node to the next one. Input layers in these models process the input information (related

to images of objects in this case), output layers produce the result of processing (a label in this case), and hidden layers that lie between these two do the work of solving the task at hand.

The most prominent class of models in the context of object recognition modelling are deep convolutional neural networks (see Fig 1). A general scheme of how DCNNs solve the problem of object recognition assumes, in the first step, extracting features from the input image, such as edges, curves, colours, etc. This step is performed by iterative employment of convolutional layers, the nonlinear activation function, and max-pooling layers, all of which are biologically inspired (Kreiman, 2021; LeCun et al., 2015). These three kinds of nodes jointly enable fine-tuned detection of features of increasing complexity across the processing hierarchy, in ways that are invariant to idiosyncratic presentations of objects, thus meeting the challenges of specificity and invariance for object recognition. In the second step, a classifier in the form of fully connected layers learns to map these features to object labels.

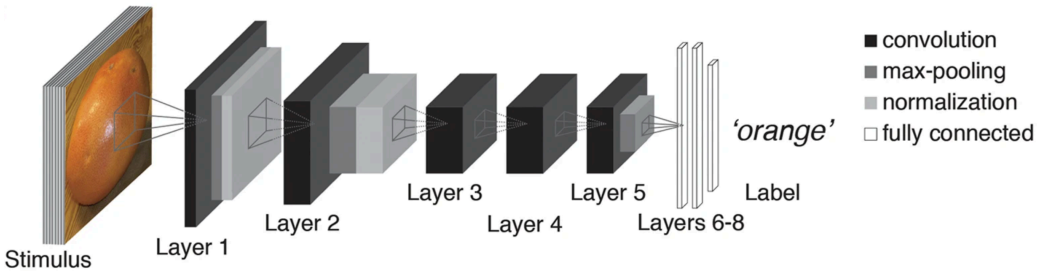

Fig 1 AlexNet, an eight-layer network, reproduced from Cichy et al. (2016). Layers 1-5 iterate convolution and max-pooling operations. Convolutional nodes are organised in feature maps, where each node in a feature map detects the same feature in its own receptive field, and they jointly tile the whole visual field. There are many feature maps per layer. Max-pooling nodes help with meeting the challenge of invariance. Extracted features from layer 5 are then passed onto layers 6-8 which are fully connected and approximate a readout from a population of neurons, in order to output a label

This is a general scheme of how DCNNs solve the core object recognition task.[1] Various architectures have been developed that exhibit these general design features, but nevertheless widely differ in their other architectural features.[2] Networks may differ regarding their depth, number of nodes, and number of connections between nodes, but also other architectural motifs. For example, many of them are feedforward networks, which seems plausible given very fast reaction times of around 350 ms in the task (DiCarlo et al., 2012), the finding that constrains the role of feedback connections for core object recognition. In contrast, some are recurrent networks, that incorporate feedback connections.

In the performance optimisation-driven approach (Yamins & DiCarlo, 2016), during the process of training the workings of DCNNs are not constrained by brain data. DCNNs are trained in a supervised manner for the object recognition task, learning on their own which features are useful for the task. Trained DCNNs enter the process of model selection using frameworks that compare their workings with neural population coding in the brain when they are exposed to the same stimuli in the task.

---

[1] For a philosophically accessible introduction see Buckner (2019).

[2] For overviews of types of architectures see Storrs et al. (2021) and Xu & Vaziri-Pashkam (2021).

3 Mechanistic explanatory potential of DCNNs

The motivation to discuss whether DCNNs can be explanatory models and if they can, in which sense, is motivated by a breadth of findings that gave rise to the deep learning revolution in neuroscience in the last ten years. The quest to advance modelling of visual processes by using DCNNs (Schrimpf et al., 2020b) started with the findings of DCNNs hierarchically corresponding to stages of processing along the ventral stream – early layers in DCNNs were found to be most similar to early processing stages in the ventral stream, while deeper layers were found to be most similar to late processing stages in the ventral stream (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Although visual areas in the brain have been extensively researched and much has been known about the way the early visual areas work, later visual areas have eluded visual neuroscientists because of their convoluted response properties. Thus, it was a surprising result when it turned out that DCNNs trained solely to successfully perform object recognition were able to predict neural response properties to an extent, including those of the later visual areas such as the inferior temporal cortex (Cichy et al., 2016; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). Correspondences based on representational similarity analysis, which quantifies second-order similarities between dissimilarities of stimuli-elicited patterns in DCNNs and the brain, have had an important role, lending support to the view that the ventral stream is a hierarchical system, processing more complex features as one goes from early to late processing stages.

With these empirical findings came a diversity of views about the roles of DCNNs in scientific practice and their explanatory potential.

Although researchers utilising DCNNs have been criticised for not explicitly reflecting on the kinds of explanations their models are aimed at (Kay, 2018), several reviews of the landscape of modelling the brain with DCNNs explicitly frame this modelling endeavour as an attempt to obtain mechanistic explanations of cognitive phenomena (Doerig et al., 2023; Lindsay, 2021; Schrimpf et al., 2020b; Kietzmann et al., 2019; Kriegeskorte, 2015). Modelling with DCNNs, according to these views, is a continuation of a widespread attempt to reach mechanistic explanations in neuroscience.

Recently some philosophical work has proceeded in that direction as well. Cao & Yamins (2021a, 2021b) attempt to counter long-standing doubts that DCNNs could be mechanistic models, by developing a particular account of mechanistic abstraction applicable to DCNNs and reviewing evidence that vindicates these models as mechanistic explanations of the ventral stream. Buckner (2018) gives a proposal for a mechanism implemented in both DCNNs and the ventral stream, building upon Stinson's analysis (2018) of the mechanistic explanatory potential of connectionist networks.

In sum, there seems to be some agreement across philosophy and neuroscience that DCNNs can be or that they already are mechanistic models of the human visual system. This paper speaks to that kind of view.

It is important, however, to note that there is an alternative view that considers DCNNs similarly to how neuroscientists consider animal models. According to this comparative model interpretation, DCNNs are treated as artificial model organisms (Scholte, 2018). DCNNs lend themselves to being tweaked in various ways. One can train them on different datasets and change their architectural features as one likes. Such malleability and accessibility of their inner workings can help the process of hypothesis generation about which model features are possibly instantiated in the brain. According to this view, current DCNNs are not

directly meant to model brain mechanisms but to be used as a platform for hypothesis generation that can subsequently help to form more elaborate and stringent models that would be intended as explanatory models of brain processes.

3.1 Mechanistic abstraction and its application to DCNNs

Mechanisms are entities and activities organised in such a way that they are responsible for the phenomenon to be explained (Illari & Williamson, 2012). Explaining a phenomenon, such as object recognition, amounts to describing a mechanism responsible for it. At the intersection of neuroscience and deep learning, mechanisms are often characterised as mechanisms that represent stimuli-related properties, thereby enabling appropriate task performance. The talk of representations has been ubiquitous in visual neuroscience since its early days (Bechtel, 2007), while deep learning is a representation-learning method (LeCun et al., 2015). The particular representations that DCNNs and the ventral stream acquire enable them to meet the challenges of invariance and specificity for object recognition, and researchers precisely aim to learn about the ventral stream representations using DCNN modelling (Lindsay, 2021; Kietzmann et al., 2019; Kriegeskorte, 2015), especially in its higher visual areas.

Thus, mechanisms for object recognition are mechanisms that operate on representations (Bechtel, 2008, 2007), carrying task-relevant content about stimuli. From this representational perspective, entities that comprise a mechanism are representations, and activities are operations that transform representations in spatio-temporally organised ways. Another perspective on the mechanism of object recognition is implementational, which focuses on the contribution of neural structures and their activities to the function of object recognition. The aspects of this dual perspective on the mechanism of object recognition are connected through the notion of the vehicle of representation (Bechtel, 2007, p. 34). Vehicles of representation are specifiable types of neural processes that carry stimulus-related representational content, needed for successful task performance.

A model of object recognition has mechanistic explanatory force if it has variables that map onto the representations, activities, and organisational properties of the brain mechanism, and if the dependencies posited between these variables map onto the causal relations between their counterparts in the target brain mechanism. This is the model-to-mechanism mapping (3M) requirement for mechanistic explanations (Craver & Kaplan, 2020; Kaplan & Craver, 2011). However, some opposition to the idea of DCNNs being mechanistic models of the brain comes from the fact that they aim to model brain processes at a higher level of abstraction than some paradigmatically mechanistic explanations, such as that of long-term potentiation (Craver, 2007) or depolarisation (Machamer et al., 2000). DCNNs abstract away from many low-level neural features, which opens up the question of their compatibility with the 3M requirement.

Recently there has been a lot of discussion on the topic of abstraction and mechanistic explanation (Stinson, 2016; Potochnik, 2015; Chirimuuta, 2014; Levy, 2014; Weiskopf, 2011), leading to a wider recognition of abstraction being characteristic of modelling and that abstract models can be mechanistic explanations (Craver & Kaplan, 2020; Boone & Piccinini, 2016). Models can abstract away from both, details that are irrelevant for object recognition, as well as some relevant details, and still be mechanistically explanatory in virtue of satisfying the 3M requirement. Abstract mechanistic explanatory models often aim to describe

just those core features of mechanisms that are the most important for the phenomenon of interest.

The account of mechanism schemata found in Machamer et al. (2000) and Darden (2002) captures this point.[3] Typically, scientists are not after describing all the details of a particular mechanism but aim to characterise a type of mechanism responsible for the phenomenon. A mechanism schema is an abstract description of a type of mechanism (Darden, 2002; Machamer et al., 2000). Schemas may have a general scope, capturing shared major features of mechanisms that can occur within quite different systems (Boone & Piccinini, 2016). For example, a mechanism schema for protein synthesis is DNA→RNA→Protein (Darden, 2002; Machamer et al., 2000). This abstract mechanistic template can be instantiated by inserting more detailed characterisations of entities, activities, and organisational properties for the variables in the schema, reaching a more detailed description of an instance of a mechanism for protein synthesis. Schema instantiation can play an important role in mechanism discovery (Darden, 2002). For example, if a type of mechanism in a DCNN can be depicted as a mechanism schema, it can then be transferred as a hypothesis onto the ventral stream in an attempt to see whether and how that schema is instantiated in more detail.

Deep neural network modelling may thus be shown to be continuous with other abstract mechanistic modelling efforts found in biology more broadly if it can provide abstract descriptions of the types of mechanisms responsible for phenomena to be explained. I take the core features of the mechanism type responsible for object recognition to be the representations that DCNNs and the target instantiate, cohering with the goal of the field to learn about representations in the ventral stream (Lindsay, 2021; Kietzmann et al., 2019; Kriegeskorte, 2015). The emphasis on similarities of mechanistic entities, in this case, representations, is one way to type mechanisms (Glennan & Illari, 2017).[4] These core features of a mechanism type are the ones that are taken to be key contributors to the runnability of the model and successful task performance. The runnability of the model is crucial for this modelling paradigm (Rumelhart et al., 1986) since it embeds the possibility of a model exhibiting similar object recognition task performance as humans in virtue of operating over similar representations (Cao & Yamins, 2021a; Kriegeskorte & Douglas, 2018; Kriegeskorte, 2015).

A mechanism schema of a DCNN has explanatory force if there is a mapping between the major features of a mechanism type it posits and the target system that is relevant for object recognition, as per the 3M requirement. I discuss contexts in which the mapping between DCNN and the ventral stream representations is formed using the framework of representational similarity analysis, which I turn to in section 4.

3.2 Two candidate levels for the mapping of representations

Establishing a mapping between representations in DCNNs and the ventral stream presupposes a target level of analysis relevant for object recognition. While generally different

[3] Note that Craver (2007) characterises schemata differently, as lying between more gappy mechanistic explanations (mechanism sketches) and complete mechanistic explanations.

[4] Typing mechanisms according to their representations is reflected in the scientific practice itself. For example, Storrs et al. (2021) train architecturally different types of DCNNs. When they are shown to equally correspond to the ventral stream using representational similarity analysis, the authors conclude that DCNNs develop similar representations, and that the architectural features responsible for that result are those that are shared across models. Thus, the architectural differences between models are not seen as relevant differences.

neural network models may bottom out at different levels of analysis, looking into the current research practice of using DCNNs to model the object recognition mechanism enables us to simplify the issue by considering two candidate levels.

According to one view, relevant representations lie on the level of individual neurons. Extending the traditional way of trying to understand the visual system by analysing which stimulus features individual neurons are responsive to (Kriegeskorte & Wei, 2021), this view considers individual neuronal activations as vehicles of representational content (Poldrack, 2021, pp. 1314-1315). These descriptions of individual neuronal activations are summarised in the form of tuning functions, which characterise the dependence of neuronal firing on a stimulus feature (see Fig 2). For example, the dependence of the firing of neurons in the primary visual cortex on the presence of edge-shaped stimuli has been classically characterised by a bell-shaped tuning curve. Call this the Tuning Functions account. The Tuning Functions account captures an attempt to map a mechanism schema of a DCNN on the level of tuning functions to tuning functions in a neuronal population in the brain.

An alternative view considers neural manifolds as representational vehicles. Neural manifolds are responses of a population of neurons or nodes to a variety of images of an object – displaying it from different perspectives, varying in scale, location, etc., forming a continuous, low-dimensional surface inside the high-dimensional neural representational space (DiCarlo & Cox, 2007). See Fig 2. It has been proposed that the object recognition capacity may depend on reformatting neural manifolds in representational spaces in order to make them less entangled and more separable down the processing hierarchy (DiCarlo et al., 2012; DiCarlo & Cox, 2007). Call this the Neural Manifolds account. According to it, the mechanism schema of a DCNN lies on the level of neural manifolds that one tries to map in the brain as well (Poldrack, 2021; Buckner, 2018).[5]

Mapping tuning functions and neural manifolds coheres well with the goals of the neural network-based research programme to bridge "between explanatory levels, from single units, through collective dynamics and onto behaviour" (Doerig et al., 2023, p. 436). However, whether these two accounts are compatible or competitors as explainers has been a topic of recent discussions (Barack & Krakauer, 2021).[6] Without taking a stance on this question, I charitably consider that mapping representations on both of these levels of analysis can yield an explanatory mechanism schema of object recognition.

---

[5] The epistemic status of neural manifolds is, however, under-theorised, with an exception of Humphries (2021). Various views can be found across the discourse. Some see them as descriptive, summarising tools (Whiteway & Butts, 2019; Williamson et al., 2019), others as important organisational principles of neural computation (Barrett et al., 2019). Yet another view is that they are a type of an entity the brain and DCNNs use to compute (Jazayeri & Ostojic, 2021; Vyas et al., 2020; Gallego et al., 2017).

[6] One relevant factor is that sets of different tuning functions can give rise to similarly shaped manifolds (Kriegeskorte & Wei, 2021). Thus, if one's goal is to obtain a description of neural tuning, bottoming out on the level of manifolds will not be satisfying.

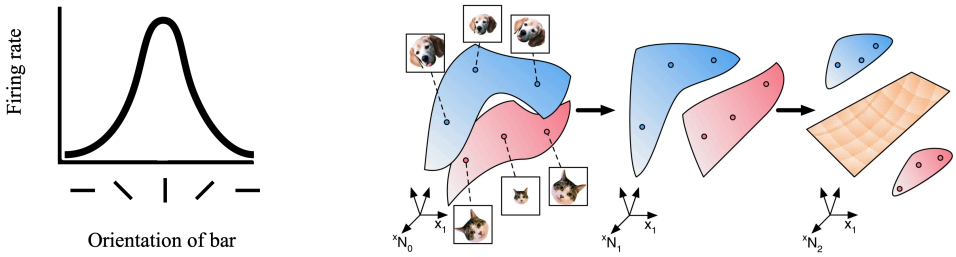

Fig 2 Tuning functions and neural manifolds. Left: A neuronal tuning function. Right: Neural manifolds across three processing stages, reproduced from Cohen et al. (2020). Notice the change of the shape of manifolds as they become more disentangled

4 Comparing representational geometries of DCNNs and the ventral stream

Representational similarity analysis (RSA) is a widely utilised framework in cognitive neuroscience for assessing the similarity of processing of systems (Kriegeskorte et al., 2008). Taking aside some recent intervention-based ways to compare the workings of DCNNs and the brain (Sexton & Love, 2022; Bashivan et al., 2019), RSA has been one of the two most frequently used ways of comparing representations of brains and DCNNs, besides linear mapping of neuronal activations based on activations of nodes in a DCNN (Schrimpf et al., 2020a). Thus, it forms an important evidential basis for the assessment of the mechanistic explanatory potential of DCNNs. Indeed, several scientific reviews mention RSA as a method of comparing DCNN with brain representations with the goal of reaching mechanistic explanations (Doerig et al., 2023; Lindsay, 2021; Kietzmann et al., 2019; Kriegeskorte, 2015).

In the philosophical literature, Buckner (2018) and Cao & Yamins (2021a) argue for the mechanistic explanatory potential of DCNNs but they do not analyse whether RSA-based evidence can corroborate DCNNs toward that goal. An exception to this is the work of Kieval (2022) who argues that RSA can help uncover brain mechanisms via shared causal patterns between DCNNs and the brain.

RSA is an analysis method that allows for an examination of a large number of stimuli-elicited patterns, obtained by fMRI in human participants and by recording activations of nodes in a DCNN. It is a pattern analysis method that analyses activations of DCNNs and brains multivariately as a population code. The activity of a population of neurons (or nodes in a layer) elicited by a presentation of a stimulus is jointly taken as representing properties of that stimulus.

Representations across voxels in a brain region or across nodes in a DCNN layer are depicted in a multidimensional space, where each dimension of the space stands for the activation level of a voxel in a brain, or a node in a DCNN. A pattern of activation elicited by a presentation of a visual stimulus across a population of voxels or nodes is then represented as a point in this multidimensional representational space. Thus, all the stimuli presented to a brain region or a DCNN layer have their corresponding points in this multidimensional space.

RSA works by quantifying the similarity structure of activation patterns elicited by a set of stimuli across stages of processing, also called representational geometry (Kriegeskorte et al., 2008a). Importantly, RSA does not aim to directly correspond architectures of systems, numbers of nodes or neurons in a population, and similar – these may be substantially different across systems that are compared. The framework aims to compare representational geometries across systems, which are of course dependent upon the particularities of the architectures they have.

To characterise a representational geometry for a brain region in a task (or a layer in a DCNN), one forms a representational dissimilarity matrix (RDM). This is the first step of the RSA methodology. In the second step, the similarity of RDMs of a brain region and a DCNN layer is assessed.

In the first step, one forms an RDM for a brain region by quantifying how dissimilar each two stimuli-elicited patterns are. Another RDM is formed for a DCNN layer. For example, if the stimuli set are images of animate objects, such as faces, and inanimate objects such as houses, then one calculates the dissimilarity between every two patterns elicited by animates and inanimates in that set. An RDM is a square matrix (see Fig 3), whose rows and columns are indexed by visual stimulus presentations, in this case, images of animate and inanimate objects. Then, each cell in the matrix contains a numerical value standing for the measured dissimilarity between two stimuli-elicited patterns. The matrix is symmetric about a diagonal of zeros (standing for the dissimilarity of each pattern from itself).

To quantify dissimilarity between two stimuli-elicited patterns one looks at their distance in the multidimensional space. The bigger the distance, according to a particular similarity measure, the bigger the dissimilarity between the patterns. Often this measure is correlation (Kriegeskorte et al., 2008a), which calculates distances as 1−correlation across voxels in the brain, or nodes in a DCNN (1 standing for perfect correlation minus their actual correlation). Thus, intuitively, an RDM captures how dissimilarly some images in a set of stimuli are processed by a model or the target system.

For example, Fig. 3 illustrates the responses of the inferior temporal (IT) cortex in humans clustering into two broad categories of animate and inanimate objects. IT representations of animate objects are more similar to each other than to IT representations of inanimate objects, and vice versa. To the extent a new incoming stimulus-elicited pattern is similar to patterns elicited by animates, it will tend to have similar effects on downstream neurons and the object recognition performance as other animate objects. Such analysis of activation patterns through the lens of similarity suggests which representational distinctions the IT cortex makes in order to successfully perform the task.

After the formation of RDMs of a DCNN layer and a brain region, in the second step, their dissimilarity is assessed. That comparison often looks at how correlated their RDMs are and quantifies their dissimilarity as 1−correlation as well.

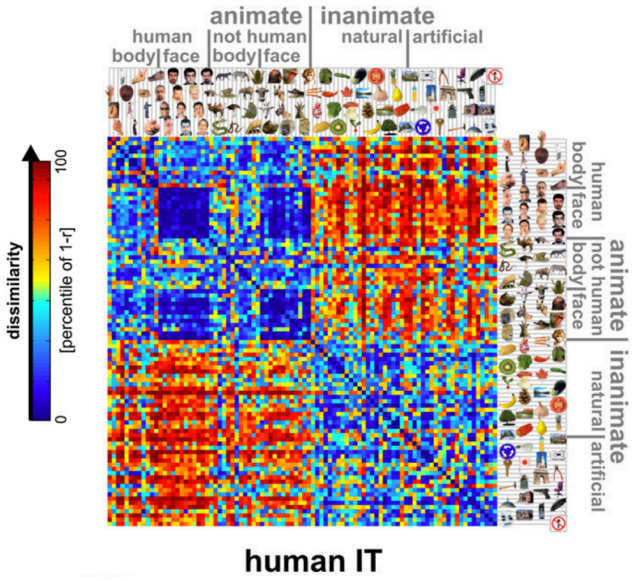

Fig 3 An example of an empirical RDM obtained in the human IT cortex, reproduced from Kriegeskorte et al. (2008b). The most noticeable clusters are those formed by low dissimilarities in the upper left quadrant between patterns elicited by animate objects, and lower right quadrant for patterns elicited by inanimate objects. The other two quadrants describe high dissimilarities between patterns elicited by categories of animates and inanimates

4.1 Quantifying Dissimilarity

The formation of an RDM requires the adoption of a similarity measure. While cognitive psychologists extensively discussed the notion of similarity and the extent to which it can be invoked as an explanatory or an evidential construct (Edelman, 1999; Medin et al., 1993), such efforts are not paralleled in contemporary neuroscience. Taking a look over the current research practices of using RSA to compare brains and DCNNs, or DCNNs among themselves, reveals that the field is not unified in quantifying similarity in a particular way. Although correlation is often used as a similarity measure, other similarity measures are also in play. Similarity measures used are correlation and cosine distance (Xu & Vaziri-Pashkam, 2021; Mehrer et al., 2020; Cichy et al., 2016; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014), Euclidean and Mahalanobis distances (Storrs et al., 2021; Xu & Vaziri-Pashkam, 2021; Mehrer et al., 2020), while Kornblith et al. (2019) devise a way of quantifying representational geometry based on the dot product.

I contrast members of two broad families of similarity measures – correlation and cosine distance, on the one hand, and Euclidean and Mahalanobis distances, on the other (Bobadilla-Suarez et al., 2020). I primarily focus on correlation and Euclidean distance to illustrate the fact that similarity measures provide different perspectives on what distance is in representational space. Then in section 5 I argue that correlation and Euclidean distance pick out different types of mechanisms according to both the Tuning Functions account and the Neural Manifolds account.

Correlation and cosine distance are angle-based measures (Bobadilla-Suarez et al., 2020). Take as an example a multivariate space spanned by the activations of two voxels to an image of an orange and an image of a pear (see Fig 4). Then one can imagine looking in the direction of the two stimulus-elicited patterns A and B from the origin of this two-dimen-

sional space. According to correlation and cosine distance, the distance between these two stimulus-elicited patterns is a function of the angle that these two points subtend with the origin. Cosine distance measures similarity as the cosine of the angle that they subtend, while dissimilarity is calculated when that number is subtracted from 1. Geometric interpretation of correlation relies on the same logic, so that the distance between the two vectors is the cosine of the angle they subtend after mean centering each condition (Ramírez, 2018; Walther et al., 2016; Mur et al., 2009).

In contrast, Euclidean and Mahalanobis distance measures are magnitude-based (Bobadilla-Suarez et al., 2020). While Mahalanobis distance measures the distance between a point and a distribution, Euclidean distance is a non-distributional measure. Turning again to Fig 4, the distance between two stimulus-elicited patterns A and B according to Euclidean distance is their distance in Euclidean space (dotted lines in Fig 4). Intuitively, in the case of Euclidean distance in this example, one looks at the space from above and measures distances that way. Euclidean distance between two vectors is calculated as the square root of the summed squared differences along each vector component (Ramírez, 2018; Walther et al., 2016; Mur et al., 2009).

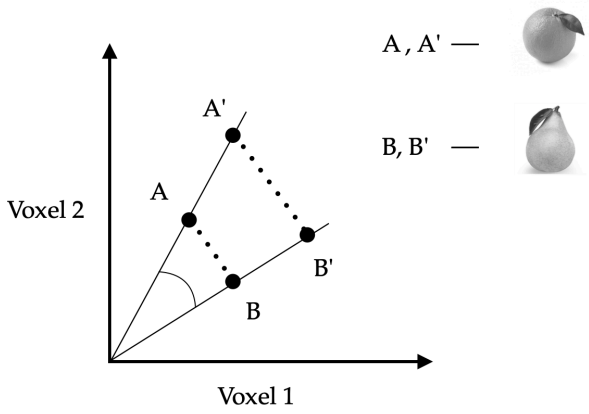

Fig 4 Correlation and cosine distance measures are angle-based measures – similarity between an orange-elicited response pattern and a pear-elicited response pattern is a function of the angle they subtend with the origin, and stays the same across all pairs of vectors (A & B, A' & B') subtending the same angle. In contrast, Euclidean distance is a magnitude-based measure that relies on measuring distances as the length of a straight line so that the distances between vectors (dotted lines) are different

Correlation and Euclidean distance provide different respects in which one can quantify representational geometries. Since correlation relies on measuring angle-based distances after the mean value subtraction, the overall mean in the region is disregarded as a coding dimension (Ramírez, 2018; Ramírez et al., 2014). Thus, correlation measures distances by looking at the pattern of differential population firing across conditions. In direct contrast to this, as a magnitude-based measure Euclidean distance quantifies representational geometries

in virtue of tracking differences in amplitude of population firing, and consequently the mean is taken as a coding dimension (Ramírez, 2018; Ramírez et al., 2014).[7]

Let us go back to the categorical structure of IT representations in Fig 3. The RDM reveals two clusters corresponding to animates and inanimates. Correlation, then, implies that the between-category information needed for object categorisation is not carried by the overall amplitude of the population response. Consequently, a downstream neural population does not discriminate between patterns elicited by animates and inanimates in virtue of the overall amplitude, which, in turn, implies that outputting an appropriate label in the object recognition task does not depend upon it either. In contrast, Euclidean distance sees precisely the differences in amplitude as properties of population activations that carry such discriminative information and lead to successful task performance.

4.2 Invariance properties of similarity measures

As mentioned in section 4, the basis of RSA is the representation of stimuli-elicited patterns in a multidimensional space. However, that representation abstracts away from tuning functions of neurons in voxels or nodes in a DCNN, focusing on the population-level response. Nevertheless, the overall population response in a voxel or a DCNN layer is of course dependent upon tuning functions of individual neurons or nodes that comprise them. In the case of RSA, a set of neurons in a population or nodes in a layer tuned for some stimuli-related properties fully determines a corresponding representational geometry (Kriegeskorte & Wei, 2021).

However, the same representational geometry can be implemented by different sets of tuning functions (Kriegeskorte & Wei, 2021). This is a consequence of the invariance properties of similarity measures.

Invariance properties determine the conditions under which perturbations of points in the representational space conserve a representational geometry quantified using a particular similarity measure (Kornblith et al., 2019). Perturbations that a similarity measure is invariant to are not considered a relevant transformation of vectors for the quantification of similarity. In practice, this means that a DCNN layer could exhibit a pattern of responses that is some transformation of responses of human IT, but a transformation that does not affect much its representational geometry.

For example, correlation is scale invariant (see Fig 4). Uniformly multiplying vectors in state space with a certain number does not affect how similar stimuli-elicited patterns are, as long as they keep subtending the same angle. From the perspective of the origin, one cannot detect how far away A and B are from A' and B', since both pairs of vectors lie on the same horizon. In contrast, being a magnitude-based measure Euclidean distance considers differences in amplitude as meaningful differences. Perturbations affecting the amplitude affect representational geometries.

Another example is the rotation of points in multivariate space. In Fig 5, after rigidly rotating vectors A and B in space along with other data points, their similarity will stay the same according to Euclidean distance (Mehrer et al., 2020). In contrast, correlation is not in-

[7] See Walther et al. (2016, Fig 1) for other illustrations of the ways Euclidean-based and correlation-based geometries can be affected by common data manipulation techniques.

variant to such rotations since they affect the angle between vectors and therefore their similarities (Mehrer et al., 2020).

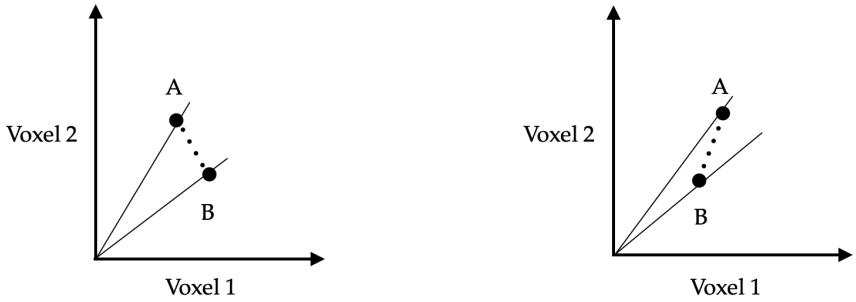

Fig 5 Rotating A and B in space affects the angle between them, which in turn affects judgments of similarity based on correlation. However, their Euclidean distance stays the same. The figure is based on Mehrer et al. (2020, Supplementary Figure 5)

Each such transformation of vectors that leaves a correlation-based or a Euclidean-based representational geometry intact often entails a change on the level of tuning functions able to implement the geometry. The same population level response and its representational geometry can be yielded by populations of neurons with different tuning functions (Kriegeskorte & Wei, 2021). That the relationship between sets of tuning functions and a representational geometry is many-to-one is a consequence of the invariance properties of similarity measures.

4.3 The choice of similarity measure is theoretically non-trivial – a case study

In the previous two sections, I have discussed how similarity measures pick out different properties of population level responses in order to quantify a representational geometry, and that due to the invariance properties of similarity measures multiple sets of tuning functions are able to implement the same representational geometry. In this section, I illustrate that correlation-based and Euclidean-based representational geometries are implemented by non-identical sets of tuning functions. The choice of a similarity measure one applies is non-trivial since it can lead to theoretically meaningfully different conclusions.

To illustrate my point I turn to the study of Ramirez et al. (2014), which deals with the representations of faces at different rotational angles in human fusiform face area (FFA). Responses in FFA were recorded using fMRI. While participants were fixating on the centre of the screen, stimuli of faces at different angles (−90°, −45°, 0°, 45°, 90°) were shown at two positions – above and below the fixation cross.

Ramirez et al. quantify the representational geometry of faces at all rotational angles in FFA, for two positions (above and below the fixation cross). See Fig 6. They tested two hypotheses related to the kind of tuning that can account for the representational geometry in FFA. One hypothesis was that neurons in FFA are primarily unimodally tuned for a single preferred view. According to another hypothesis, FFA neurons could exhibit mirror-symmetric encoding of face orientations, so that the responses are highly similar for faces presented at −45° and 45°, as well as −90° and 90°. Two representational distance matrices were

formed capturing which representational geometry of a region would be expected if it were the case that unimodally tuned neurons encode face orientation across retinal positions, and if it were the case that mirror-symmetric neurons encode it. Finally, these two representational distance matrices were compared with the empirical distance matrix of the FFA responses.

When the authors used correlation to quantify the representational geometries, the unimodal tuning hypothesis was favoured, as it captured more variance than the geometry based on mirror-symmetric tuning. However, when Euclidean distance was used, the two hypotheses were indistinguishable.

Therefore, the choice of similarity measure used to quantify the representational geometry of FFA has a consequence for a theoretically meaningful conclusion that is drawn. Unimodal tuning is a hypothesis that stems from view-dependent theories of object and face representation, and it is a hypothesis that is favoured if correlation is chosen to quantify the representational geometry of FFA. On the other hand, the hypothesis of mirror-symmetric tuning was first put forth by proponents of view-independent theories of the object and face representation according to which individual cells respond to mirror-symmetric rotational angles, thus partially achieving view-point invariance (Ramírez, 2018). However, when representational geometries were quantified with Euclidean distance, both unimodal as well as mirror-symmetric tuning could account for the representational geometry of human FFA.

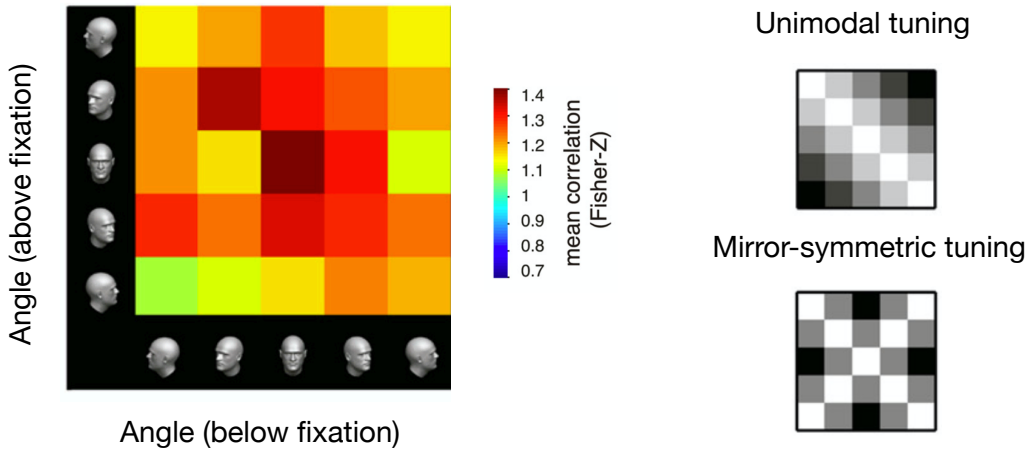

Fig 6 An empirically obtained correlation-based RDM of face rotational angles in FFA across two retinal positions (left), adapted from Ramirez et al. (2014). If correlation is used to quantify RDMs, the unimodal tuning hypothesis better accounts for the RDM. If Euclidean distance is used, both hypotheses can account for the RDM.

5 Implementational view on the mechanism of object recognition: different similarity measures correspond different types of mechanisms

When RSA provides the mapping between DCNNs and the brain, in order to satisfy the 3M requirement, an issue arises due to a variety of similarity measures used in that framework. In this section, using section 4, I argue that correlation and Euclidean distance pick out different representational vehicles, according to the Tuning Functions and the Neural Manifolds accounts, that they correspond in DCNNs and the brain.

If one types mechanisms according to their representations (see section 3.1), the implementational view on the mechanism of object recognition allows for typing mechanisms according to the representational vehicles systems have. Representational vehicles are types of neural states that carry representational content, and that are causally relevant for the operation of a mechanism downstream and its task performance (Bechtel, 2007). If it is shown that different similarity measures pick out different representational vehicles, then they pick out different mechanism types that they correspond in DCNNs and the ventral stream.

5.1 The Neural Manifolds account

It may seem far-fetched at first sight that the quantification of representational geometries can in principle capture something about the underlying mechanism of the system, since analyses of similarities of processing may seem too detached from the actual workings of the mechanism. This, however, is incorrect. Representational geometries can in principle reveal some core features of representational vehicles responsible for object recognition.

Let us go back to the example of the representational geometry in human IT in Fig 3, which is reproduced by a DCNN (Khaligh-Razavi & Kriegeskorte, 2014). The structure of similarities between responses elicited by images of animate and inanimate objects form two clusters. These low-dimensional subspaces revealed in IT and a DCNN layer are neural manifolds of animates and inanimates. The structure of similarities between patterns elicited by animates and inanimates describes the shape of their neural manifolds. This may be more visible in Fig 2, where the shape of cat and dog manifolds is depicted. A quantification of representational geometry can capture the shapes of representational vehicles according to the Neural Manifolds account.

Shapes of manifolds have a potential causal relevance for processing downstream. From the perspective of a downstream neural population or a DCNN layer that is meant to distinguish between animates and inanimates and output a label in the task, similarity relations between stimuli-elicited patterns lying on different manifolds serve to distinguish them (Kriegeskorte & Wei, 2021; Buckner, 2018; Kriegeskorte & Kievit, 2013). For example, if a newly incoming stimulus-elicited pattern is more similar to the previously encountered patterns elicited by animates and thus falls into the manifold for animates, it will tend to have similar effects on downstream neurons and the object recognition task performance as other animate objects – it will get recognised as an animate object. This similarity structure of inputs is implicit in the weights of a DCNN – the weights ensure the treatment of new but similar inputs similarly. The generalisation ability may thus depend on the sensitivity of DCNNs and the ventral stream to the similarities between previously encountered inputs and the new ones (Rumelhart et al., 1986).

The quantification of representational geometry can describe shapes of neural manifolds relevant for processing downstream and the object recognition task performance. However, the issue arises that the shape of manifolds is defined differently depending on whether correlation or Euclidean distance is used to quantify a corresponding representational geometry. As discussed in section 4.1, correlation does not describe the shape of neural manifolds in terms of the amplitude, while Euclidean distance does precisely that. Correlation thus does not take the amplitude of a population level response as defining of representational vehicles, in virtue of which a downstream population can distinguish between animates and inanimates

and output a label in the task. The application of Euclidean distance entails the exact opposite conclusion.

Thus, in principle a representational geometry can capture shapes of manifolds as representational vehicles that are defining of a mechanism type. The issue is, however, that different similarity measures pick out different types of mechanisms in order to correspond DCNNs and the ventral stream.

5.2 The Tuning Functions account

The categorical structure of the population code in Fig 3, exhibited by a DCNN as well (Khaligh-Razavi & Kriegeskorte, 2014), is derivative from the activations of many individual neural cells or nodes. Even though it may seem that representational geometries are too detached from tuning functions, they nevertheless directly depend upon them (Kriegeskorte & Wei, 2021). A representational geometry can in principle capture some abstract properties of similarity of responses arising from a population tuned in a particular way. These properties of similarity determine which sets of tuning functions can implement a given geometry. However, it is important to note two things based on the discussion in section 4.

First, due to the invariance properties of similarity measures, as discussed in section 4.2, many sets of tuning functions can exhibit the same representational geometry. Right from the start, the relationship between tuning functions and a representational geometry is many-to-one. Thus, one should not conclude that two systems have similar tuning functions in case they have similar representational geometries.

Second, as discussed in section 4.3 in the case of Ramirez et al. (2014) study, sets of tuning functions that can implement a correlation-based representational geometry and a Euclidean distance-based representational geometry are non-identical, a matter that can be fairly theoretically meaningful as their example shows. Quantifying geometry in a particular way does fix a set of tuning functions that can implement it, as not all tuning functions will do. However, different similarity measures pick out non-identical sets of tuning functions.

Therefore, depending on how one chooses to quantify a representational geometry, different sets of tuning functions will seem to be plausibly instantiating it. If tuning functions are representational vehicles, different similarity measures pick out different sets of representational vehicles. Since the implementational view on the mechanism of object recognition types mechanisms according to their representational vehicles, different similarity measures pick out different mechanism types to correspond across DCNNs and the ventral stream.

6 Representational view on the mechanism of object recognition: RSA does not correspond representational content across systems

While similarities of representational geometries can in principle be mechanistically informative similarities for the implementational view on the mechanism of object recognition, as argued in the previous section, they underdetermine the similarity of representational content, relevant for the representational view on the mechanism of object recognition.

Not infrequently an inference is made that RSA precisely allows for a mapping between representational content across systems. Consider Kieval's (2022) conclusion that a match-up of representational geometries between a model and the brain occurs "precisely because they both instantiate the same causal patterns between mechanism and stimulus condi-

tions" (Kieval, 2022, p. 19). Roskies (2021, p. 5926) also states that "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". Such inferences that slide from the acknowledgment of correspondence of representational geometries to the claim of similar representations being used to track properties of stimuli also occur in the scientific literature (for some examples see Dujmović et al. (2022)). There is enough empirical evidence available showing that such inferences are fallacious.

6.1 The Neural Manifolds account

A representational geometry describes relational properties of similarity between stimuli-elicited patterns in a system. It is entirely consistent with a significant match-up of representational geometries between two systems that one system exhibits given relational properties of similarity in virtue of tracking shape-related properties of stimuli, while another exhibits given relational properties of similarity in virtue of tracking texture-related properties. In fact, humans rely on shapes in object recognition, while DCNNs often rely on texture in order to perform the task (Bowers et al., 2022). Geirhos et al. (2018) showed this for certain types of DCNNs, whose representational geometries, on the other hand, have been shown by Storrs et al. (2021) to be equally similar to that of human IT. This suggests that the possibility of two systems exhibiting similar representational geometries while nevertheless representing different properties of the world depends in part on the structure of the world. DCNNs are particularly successful in task contexts where there is some repeated structure present in the environment (e.g. based on texture), which can be exploited by utilising convolutions and weight sharing applied hierarchically, building up more and more complex features out of simpler ones. The brain could perform the same task by exploiting yet another repeated structure in the environment (e.g. related to shapes). If objects similar in texture are also similar in shape ("the mimic effect" as Dujmović et al. (2022) call it), then DCNNs and the ventral stream may exhibit similar representational geometries, while tracking different stimulus-related properties.

If this analysis is correct, then neural manifolds across DCNNs and the ventral stream may have similar shapes but implement different representational spaces – a shape-related and a texture-related representational space. The match of representational geometries underdetermines the similarity of representational content.

6.2 The Tuning Functions account

Given that such properties of manifolds are dependent upon tuning functions in the population (Kriegeskorte & Wei, 2021), DCNN nodes and the ventral stream neurons would have to be tuned to shape-related and texture-related properties of the stimuli as well.

Another example illustrating that the match of representational geometries underdetermines similarities of representational content across systems is based on the case study of Ramirez et al. (2014), discussed in section 4.3. When the representational geometry of FFA was quantified using Euclidean distance, both unimodal tuning as well as mirror-symmetric tuning were able to account for it, two competing hypotheses stemming from different theoretical frameworks on object and face recognition (Ramírez, 2018; Hummel, 2013). The former derives from view-dependent theories of object and face representations, while the latter

derives from view-independent theories. Thus neurons and nodes can be tuned to different properties of stimuli while exhibiting similar representational geometries. Additional problems arise due to different similarity measures – as the example of Ramirez et al. (2014) shows, they may lead us to ascribe different representational content to individual neurons.

Therefore, similarities of representational geometries alone underdetermine similarities of content, because the relationship of content ascriptions to vehicles instantiating similar representational geometries is many-to-one, contra Roskies (2021) and Kieval (2022). As DCNNs evolve and model selection based on RSA continues, this point should be kept in mind.

7 DCNNs are not how-actually mechanism schemata of object recognition

A mechanism schema is how-actually explanatory if there is a mapping between the core features of a mechanism type it posits and the target system that is relevant for object recognition, as per the 3M requirement (Craver & Kaplan, 2020). However, the framework of RSA comes with a diversity of similarity measures serving the role of the mapping function, which pick out different types of mechanisms on the level of representational vehicles that they map in a model and the target system, as argued in section 5. But which type of mechanism is relevant for object recognition? If there is no arbitration between them in terms of relevance for object recognition, it is clear that current DCNNs are not how-actually mechanism schemas of object recognition.

That the components of an explanatory mechanism have to be responsible for the phenomenon to be explained is a key aspect of mechanistic explanation that has been stressed by all prominent definitions of mechanisms (Illari & Williamson, 2012). However, the community of researchers using RSA to compare DCNNs and the brain relies on applying different similarity measures without any arbitration in terms of relevance for the explanandum capacity of object recognition. In relation to correlation and Euclidean distance, does the amplitude of the population level response carry discriminative categorical information a downstream area may use?

Except for a couple of exceptions (Bobadilla-Suarez et al., 2020; Ramírez, 2018; Ramírez et al., 2014), similarity measures and the issue of their relevance for the task at hand are not frequently discussed in neuroscience. The lack of consideration of relevance is apparent when one looks at the totality of studies using DCNNs to model object recognition. On the one hand, there are many papers (Storrs et al., 2021; Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014) that aim to assess the similarity of the brain and DCNN workings by focusing on the neural level of analysis. They do not analyse whether the properties similarity measures pick out to compare systems on the neural level of analysis are relevant for the phenomenon to be explained, in this case, object recognition task performance.

On the other hand, there are studies like Geirhos et al. (2020) that probe the workings of systems, in this case architecturally different DCNNs, with atypical stimuli like silhouette pictures or drawings. They then analyse whether they make similar kinds of errors on a trial-to-trial basis like humans. However, this analysis of the task performance is not related to the neural level of analysis across these systems.

The work on the neural level of analysis and on the level of task performance proceeds in parallel as things now stand and they are not brought in contact (with the exception of Ding et al. (2021), unrelated to RSA). But the work on mechanisms in philosophy of sci-

ence precisely suggests that relating quantified representational geometries to task performance is of crucial importance for the success of mechanistic explanation (see also Ritchie et al. (2019)), and the idea of runnability of DCNNs (section 3.1) embeds the same assumption.

One useful test of the relevance of a similarity measure for object recognition would be to see if it manages to predict the similarity of task performance across compared systems, based on the quantified similarity of their population codes. For example, if one took different DCNN architectures, tracked which kinds of errors they make on the level of task performance, and probed their workings with atypical stimuli similar to Geirhos et al. (2020), would their similarity of task performances correlate with the similarity of their representational geometries quantified by a particular similarity measure? If a similarity measure does track properties in virtue of which one can quantify the structure of the population code which would be predictive of the similarity of task performance across these DCNNs, then it would not happen, for example, that DCNNs are judged as highly similar on the level of geometries but turn out to be dissimilar on the level of task performance. In that case, we would say that a similarity measure likely does not track properties relevant for object recognition. Certainly, intervening on neural populations and altering the structure of the population code would be the gold standard for testing its relevance but looking into the correlation between similarity across mechanisms and across their task performance would be a good start.

Given that the relevance of similarity measures used as a part of the RSA framework is not assessed, it is underdetermined which of the mechanism types picked out by similarity measures are relevant for object recognition. Thus, DCNNs are not how-actually mechanism schemas of object recognition. While explanatory models help us answer what-if-things-had-been-different questions about a target system's behaviour, DCNNs do not do this with respect to representational vehicles.

8 DCNNs are not how-possibly mechanism schemata of object recognition

It is not just the practices surrounding the use of similarity measures that hinder reaching the goal of mechanistic explanation. The crucial part relates to the ways DCNN models are architecturally constructed, giving rise to the issues of underdetermination of similarity measures. Few and biologically implausible architectural constraints leave DCNNs too highly idealised and too detached from the target system for the purposes of explaining representational mechanisms in the brain, even in a how-possibly way.

A how-possibly model does not demonstrate only that something is possible, in this case reproducing object recognition based on a DCNN architecture, but it accounts for how it occurs (Brainard, 2020), elucidating the link between the explanandum and the explanans (Machamer et al., 2000). In this context, this amounts to portraying some core properties of representational vehicles that define a mechanism type underlying object recognition.

A how-possibly model is a loosely constrained conjecture (Weiskopf, 2011; Craver, 2007). The amount of evidential support for it can be low. However, the value of such models is in linking a hypothesised type of mechanism to an explanandum in a way that constrains the domain of models one may want to explore further by introducing a boundary condition on the space of possible models.

Take a classic example of a how-possibly model – Schelling's checkerboard model of residential segregation (1971). The model is highly idealised but posits one core difference-

maker for the explanandum – that people do not prefer the minority status. While it was previously thought that only racial discrimination could lead to residential segregation, this model transformed our understanding of the explanandum by showing that it can be a result of non-discriminatory preferences. In order to fulfil the role of a how-possibly model, this model posits a difference-maker that acts as a boundary on the class of models we may want to explore further.

Current DCNNs are not analogous. In its interaction with the application of correlation and Euclidean distances, one and the same DCNN can be seen as instantiating two different mechanism schemas, that imply contradictory answers to what-if-things-had-been-different-questions, for example about the role of the overall response amplitude as the vehicle of representation. Is the class of models researchers should further investigate such that models instantiate a mechanism schema according to which the overall amplitude of the population response matters for representational vehicles, or the one according to which it does not matter? Approaching from the angle of tuning functions – which mechanism schema outlines a promising class of models regarding tuning across stages of processing? Current DCNNs do not delimit the space of promising models across similarity measures-induced mechanism schemas of their workings. They do not play that directive role expected of a how-possibly model, by helping us navigate the space of the population coding strategies on the level of manifolds or tuning. Since DCNNs do not impose such a boundary, they are not how-possibly mechanism schemata of object recognition. We are left in the dark about the representational and implementational properties of the mechanism of object recognition – they may lie anywhere in the vast space of vehicles and content that can yield representational geometries quantified in diverse ways.

This rampant underdetermination happens because of the highly idealised status of DCNNs. The types of architectural constraints embedded in current models often do not reflect theoretically informed hypotheses on constraints on population coding strategies the brain may use. Many current architectures within the performance optimisation-driven approach embed only several architectural constraints (Storrs et al., 2021; Xu & Vaziri-Pashkam, 2021), some of which are known to be implausible. For example, many DCNNs aim to mimic the hierarchical processing of the visual cortex in a feedforward way although the ventral stream is a seat of a lot of recurrent processing, and the number of layers may not meaningfully correspond to the stages of processing in the ventral stream (models may have from 8 to 200 layers). Some features are invoked for engineering reasons rather than the reasons of biological resemblance, such as inception modules or residual connections.[8] These architectural features do not allow DCNNs to converge to solutions that would be sufficiently constraining and instructive for the goals of learning about representations in the target system. If DCNNs were made in theoretically more informed ways about the biological structure of the ventral stream (Revsine et al., 2023), this would help them converge to solutions that would not be as under-constraining for our goals in their interaction with RSA and would be able to impose a boundary on the class of possible representational vehicles.

Contrast this current performance optimisation-driven approach with some classical models such as, for example, a model proposed by Riesenhuber & Poggio (2000). The model expanded on HMAX (Riesenhuber & Poggio, 1999) which accomplished size and translation

[8] This approach can be contrasted with that of Revsine et al. (2023) who offer a how-possibly neural network model reflecting biologically inspired architectural features, that in interaction with stimuli properties can explain RSA-based discordant results about tuning properties in higher visual areas.

invariance by positing cells tuned to specific views of specific objects with an addition in the form of a hypothesis about object-tuned cells ensuing after view-tuned cells, where the former accomplish invariance to rotation in depth and illumination by pooling over the latter. In the case of this model, there is an explicit hypothesis about how representations differ across stages of processing, and how they accomplish invariance necessary for object recognition. Comparing such a model using RSA with the brain is much less underdetermining than in the case of DCNNs.

The architectural constraints of current DCNNs do not reflect similar theoretical considerations about the biological structure and representational strategies across stages of processing. DCNNs are then compared with the brain using a variety of similarity measures that are not arbitrated between on the grounds of relevance. Hence, this whole practice of DCNN model building and RSA-based comparisons proceeds *entirely* in a theory-free way about population coding. Consequently, object recognition is not mechanistically explained either in a how-possibly or how-actually way.

9 Conclusion

It has been claimed both across neuroscience and philosophy that DCNNs can be or that they already are mechanistic explanations of object recognition, while RSA has been taken to be able to corroborate DCNNs as mechanistic explanations. The arguments presented suggest that there are aspects of the current scientific practice of using DCNNs with RSA that are not conducive to the goal of mechanistically explaining object recognition and learning about the representations in the ventral stream. There are readily available constraints the scientific practice could use that would be beneficial towards this goal. The first suggestion is to architecturally constrain DCNNs more carefully, by invoking biological constraints known to exist in the ventral stream. More theoretically informed architectural constraints would help models converge to solutions that would not be under-constraining for the purposes of learning about representational vehicles in the target system and would be able to reduce the underdetermination arising from multiple similarity measures used in this context. The second suggestion is that the relevance of similarity measures should also be tested rigorously. Given that current scientific practice does not reflect these concerns, DCNNs are neither how-actually nor how-possibly mechanistic explanations of object recognition.

Bibliography

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359-371. https://doi.org/10.1038/s41583-021-00448-6

Barrett, D. G. T., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, *55*, 55-64. https://doi.org/10.1016/j.conb.2019.01.007

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bechtel, W. (2007). Mental mechanisms: Philosophical perspectives on cognitive neuroscience. Psychology Press. https://doi.org/10.4324/9780203810095

Bechtel, W. (2008). Mechanisms in Cognitive Psychology: What Are the Operations? *Philosophy of Science*, *75*(5), 983-994. https://doi.org/10.1086/594540

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2020). Measures of Neural Similarity. *Computational Brain & Behavior*, *3*(4), 369-383. https://doi.org/10.1007/s42113-019-00068-5

Boone, W., & Piccinini, G. (2016). Mechanistic Abstraction. *Philosophy of Science*, *83*(5), 686-697.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *Behavioral and Brain Sciences*, 1-74. https://doi.org/10.1017/S0140525X22002813

Bracci, S., & Op de Beeck, H. P. (2022). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-032720-041031

Brainard, L. (2020). How to Explain How-Possibly. *Philosophers' Imprint*, *20*(13), 1-23.

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese*, *195*(12), 5339-5372. https://doi.org/10.1007/s11229-018-01949-1

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625. https://doi.org/10.1111/phc3.12625

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Cao, R., & Yamins, D. (2021a). Explanatory models in neuroscience: Part 1--taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*. https://doi.org/10.48550/arXiv.2104.01490

Cao, R., & Yamins, D. (2021b). Explanatory models in neuroscience: Part 2--constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*. https://doi.org/10.48550/arXiv.2104.01489

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127-153. https://doi.org/10.1007/s11229-013-0369-y

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305-317. https://doi.org/10.1016/j.tics.2019.01.009

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, *11*(1), 746. https://doi.org/10.1038/s41467-020-14578-5

Craver, C. F. (2007). Explaining the brain: mechanisms and the mosaic unity of neuroscience. Clarendon Press.

Craver, C. F., & Kaplan, D. M. (2020). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *British Journal for the Philosophy of Science*, *71*(1), 287-319. https://doi.org/10.1093/bjps/axy015

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Proceedings of the Philosophy of Science Association*, *2002*(3), S354-S365.

Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., & Li, F.-F. (2009, 20-25 June 2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition.
DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333-341. https://doi.org/10.1016/j.tics.2007.06.010
DiCarlo, James J., Zoccolan, D., & Rust, Nicole C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415-434. https://doi.org/10.1016/j.neuron.2012.01.010
Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding representation similarity with statistical testing. *arXiv preprint arXiv:2108.01661*.
Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431-450. https://doi.org/10.1038/s41583-023-00705-w
Dujmović, M., Bowers, J., Adolfi, F., & Malhotra, G. (2022). The pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*. https://doi.org/10.1101/2022.04.05.487135
Edelman, S. (1999). *Representation and recognition in vision*. The MIT Press. https://doi.org/10.7551/mitpress/5890.001.0001
Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978-984. https://doi.org/10.1016/j.neuron.2017.05.025
Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890-13902.
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*. https://doi.org/10.48550/arXiv.1811.12231
Glennan, S., & Illari, P. (2017). Varieties of mechanisms. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 91-103). Routledge.
Hummel, J. E. (2013). Object recognition. *Oxford handbook of cognitive psychology*, *810*, 32-46. https://doi.org/10.1093/oxfordhb/9780195376746.013.0003
Humphries, M. D. (2021). Strong and weak principles of neural dimension reduction. *Neurons, Behavior, Data analysis, and Theory*, *5*(2), 1-28. https://doi.org/10.51628/001c.24619
Illari, P., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135. https://doi.org/10.1007/s13194-011-0038-2
Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, *70*, 113-120. https://doi.org/10.1016/j.conb.2021.08.002
Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*. *Philosophy of Science*, *78*(4), 601-627. https://doi.org/10.1086/661755
Kay, K. (2018). Principles for models of neural information processing. *NeuroImage*, *180*(Pt A). https://doi.org/10.1016/j.neuroimage.2017.08.016
Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915
Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In: Oxford University Press.
Kieval, P. H. (2022). Mapping representational mechanisms with deep neural networks. *Synthese*, *200*(3), 1-25. https://doi.org/10.1007/s11229-022-03694-y
Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. International Conference on Machine Learning. PMLR.
Kreiman, G. (2021). *Biological and Computer Vision*. Cambridge University Press. https://doi.org/10.1017/9781108649995
Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417-446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148-1160. https://doi.org/10.1038/s41593-018-0210-5
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*. https://doi.org/10.3389/neuro.06.004.2008
Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (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
Kriegeskorte, N., & Wei, X.-X. (2021). Neural tuning and representational geometry. *Nature Reviews Neuroscience*, *22*(11), 703-718. https://doi.org/10.1038/s41583-021-00502-3
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://doi.org/10.1145/3065386
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539
Levy, A. (2014). What was Hodgkin and Huxley's Achievement? *The British Journal for the Philosophy of Science*, *65*(3), 469-492. https://www.jstor.org/stable/26398392
Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017-2031. https://doi.org/10.1162/jocn_a_01544
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1-25. https://doi.org/10.1086/392759
Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278. https://doi.org/10.1037/0033-295X.100.2.254
Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725. https://doi.org/10.1038/s41467-020-19632-w
Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101-109. https://doi.org/10.1093/scan/nsn044
Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254
Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is Real-World Visual Object Recognition Hard? *PLoS Computational Biology*, *4*(1), e27. https://doi.org/10.1371/journal.pcbi.0040027
Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1), 1307-1325. https://doi.org/10.1007/s11229-020-02793-y
Potochnik, A. (2015). Causal patterns and adequate explanations. *Philosophical Studies*, *172*(5), 1163-1182.
Ramírez, F. M. (2018). Orientation Encoding and Viewpoint Invariance in Face Recognition: Inferring Neural Properties from Large-Scale Signals. *The Neuroscientist*, *24*(6), 582-608. https://doi.org/10.1177/1073858418769554
Ramírez, F. M., Cichy, R. M., Allefeld, C., & Haynes, J.-D. (2014). The Neural Code for Face Orientation in the Human Fusiform Face Area. *The Journal of Neuroscience*, *34*(36), 12155. https://doi.org/10.1523/JNEUROSCI.3156-13.2014
Revsine, C., Gonzalez-Castillo, J., Merriam, E., P., Bandettini, P., A., & Ramírez, F., M. (2023). A unifying model for discordant and concordant results in human neuroimaging studies of facial viewpoint selectivity. *bioRxiv*, 2023.2002.2008.527219. https://doi.org/10.1101/2023.02.08.527219
Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., & Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761-1770. https://doi.org/10.1038/s41593-019-0520-2

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019-1025. https://doi.org/10.1038/14819
Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3*, 1199-1204. https://doi.org/10.1038/81479
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*, *70*(2), 581-607. https://doi.org/10.1093/bjps/axx023
Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. *Synthese*, *199*(3-4), 5917-5935. https://doi.org/10.1007/s11229-021-03052-4
Rumelhart, D. E., McClelland, J. L., & PDP Research Group, C. (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations. MIT press.
Schelling, T. C. (1971). Dynamic models of segregation†. *The Journal of Mathematical Sociology*, *1*(2), 143-186. https://doi.org/10.1080/0022250X.1971.9989794
Scholte, S. H. (2018). Fantastic DNimals and where to find them. *NeuroImage*, *180*(Pt A). https://doi.org/10.1016/j.neuroimage.2017.12.077
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020a). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, 407007. https://doi.org/10.1101/407007
Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020b). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413-423. https://doi.org/10.1016/j.neuron.2020.07.040
Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219. https://doi.org/10.1126/sciadv.abm2219
Stinson, C. (2016). Mechanisms in psychology: ripping nature at its seams. *Synthese*, *193*(5).
Stinson, C. (2018). Explanation and connectionist models. In M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* (1 ed., pp. 120-133). Routledge. https://doi.org/10.4324/9781315643670-10
Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience*, 1-21. https://doi.org/10.1162/jocn_a_01755
Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annual Review of Neuroscience*, *43*(1), 249-275. https://doi.org/10.1146/annurev-neuro-092619-094115
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188-200. https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.12.012
Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313-338.
Whiteway, M. R., & Butts, D. A. (2019). The quest for interpretable models of neural population activity. *Current Opinion in Neurobiology*, *58*, 86-93. https://doi.org/10.1016/j.conb.2019.07.004
Williamson, R. C., Doiron, B., Smith, M. A., & Yu, B. M. (2019). Bridging large-scale neuronal recordings and large-scale network models using dimensionality reduction. *Current Opinion in Neurobiology*, *55*, 40-47. https://doi.org/10.1016/j.conb.2018.12.009
Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), 2065. https://doi.org/10.1038/s41467-021-22244-7
Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356-365. https://doi.org/10.1038/nn.4244
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*, *111*(23), 8619-8624. https://doi.org/10.1073/pnas.1403112111