Some pitfalls of measuring representational similarity using Representational Similarity Analysis

Marin Dujmović[1*], Jeffrey S Bowers[1], Federico Adolfi[1,2], and Gaurav Malhotra[1]

[1]*School of Psychological Science, University of Bristol, Bristol, UK*

[2]*Ernst-Strüngmann Institute for Neuroscience in Cooperation with Max-Planck Society, Frankfurt, Germany*

[*]*marin.dujmovic@bristol.ac.uk*

Abstract

A core challenge in cognitive and brain sciences is to assess whether different biological systems represent the world in a similar manner. Representational Similarity Analysis (RSA) is an innovative approach that addresses this problem by looking for a *second-order isomorphisim* in neural activation patterns. This innovation makes it easy to compare latent representations across individuals, species and computational models, and accounts for its popularity across disciplines ranging from artificial intelligence to computational neuroscience. Despite these successes, using RSA has led to difficult-to-reconcile and contradictory findings, particularly when comparing primate visual representations with deep neural networks (DNNs): even though DNNs have been shown to learn and behave in vastly different ways to humans, comparisons based on RSA have shown striking similarities in some studies. Here, we demonstrate some pitfalls of using RSA and explain how contradictory findings can arise due to false inferences about representational similarity based on RSA-scores. In a series of studies that capture increasingly plausible training and testing scenarios, we compare neural representations in computational models, primate cortex and human cortex. These studies reveal two problematic phenomena that are ubiquitous

in current research: a "mimic effect", where confounds in stimuli can lead to high RSA-scores between provably dissimilar systems, and a "modulation effect", where RSA-scores become dependent on stimuli used for testing. Since our results bear on a number of influential findings, such as comparisons made between human visual representations and those of primates and DNNs, we provide recommendations to avoid these pitfalls and sketch a way forward to a more solid science of representation in cognitive systems.

Introduction

How do other animals see the world? Do different species represent the world in a similar manner? How do the internal representations of AI systems compare with humans and animals? The traditional scientific method of probing internal representations of humans and animals (popular in both psychology and neuroscience) relates them to properties of the external world. By moving a line across the visual field of a cat, Hubel & Wisel [1] found out that neurons in the visual cortex represent edges moving in specific directions. In another Nobel-prize winning work, O'Keefe, Moser & Moser [2,3] discovered that neurons in the hippocampus and entorhinal cortex represent the location of an animal in the external world. Despite these successes it has proved difficult to relate internal representations to more complex properties of the world. Moreover, relating representations across individuals and species is challenging due to the differences in experience across individuals and differences of neural architectures across species.

These challenges have led to recent excitement around multivariate analyses methods, such as Multi-Voxel Pattern (MVP) Classification, which uses machine learning algorithms to decode neural activity [4]. MVP classification assesses whether a brain region codes for a stimulus feature by examining whether the feature can be easily decoded from

neural response patterns in the region. However, there are at least two issues with using MVP classification for comparing mental representations across individuals. Firstly, just because a stimulus feature can be easily decoded from neural response patterns in a region does not imply that downstream regions in the brain actually decode this information [5]. Different individuals (or species) may use this information in different ways and MVP classification does not provide a way of capturing this. Secondly, there are methodological limitations on mapping brain regions and neural activity patterns between individuals and species. Therefore, even if two individuals represent a visual stimulus in the same manner, a decoder trained on one individual will show a significant performance drop when applied across individuals [6].

A newer addition to multivariate analysis, Representation Similarity Analysis (RSA), is specifically designed to compare representations between different systems and overcomes some of these obstacles. RSA usually takes patterns of activity from two systems and computes how the distances between activations in one system correlate with the distances between corresponding activations in the second system (see Figure 1). Rather than compare each pattern of activation in the first system directly to the corresponding pattern of activation in the second system, it computes representational distance matrices (RDMs), a *second-order* measure of similarity that compares systems based on the relative distances between neural response patterns. This arrangement of neural response patterns in a representational space has been called a system's *representational geometry* [7]. The advantage of looking at representational geometries is that one no longer needs to match the architecture of two systems, or even the feature space of the two activity patterns (see Supplementary Information, Section A for a brief history of RSA and its philosophical origins). One could compare, for example, fMRI signals with single cell recordings, EEG traces with behavioural data, or vectors in a computer algorithm

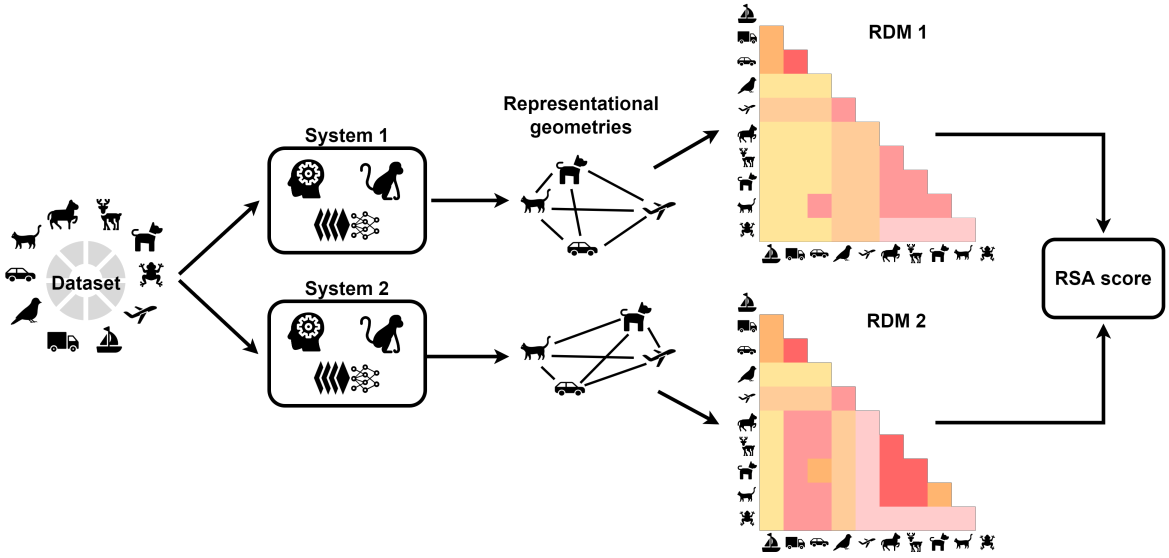

Figure 1: **RSA calculation. (A)** Stimuli from a set of categories (or conditions) are used as inputs to two different systems (for example, a human brain and a primate brain). Activity from regions of interest is recorded for each stimulus. Pair-wise distances in activity patterns are calculated to get the representational geometry of each system. This representational geometry is expressed as a representational dissimilarity matrix (RDM) for each system. Finally, an RSA score is determined by computing the correlation between the two RDMs. It is up to the resercher to make a number of choices during this process including the choice of distance measure (e.g., 1-Pearson's r, Euclidean distance etc.) and a measure for comparing RDMs (e.g., Pearson's r, Spearman's ρ, Kendall's τ, etc.).

with spiking activity of neurons [8]. RSA is now ubiquitous in computational psychology and neuroscience and has been applied to compare object representations in humans and primates [9], representations of visual scenes by different individuals [6, 10], representations of visual scenes in different parts of the brain [11], to study specific processes such as cognitive control [12] or the dynamics of object processing [13], and most recently, to relate neuronal activations in human (and primate) visual cortex with activations of units in Deep Neural Networks [14–18].

However, this flexibility in the application of RSA comes at the cost of inferences one can draw from this analysis. If the goal of the neuroscientist, psychologist or AI researcher

is to establish whether two systems are similar in mechanism, feature representation or information processing, then RSA may not be the correct analytical method to use. This is because RSA is a second-order measure – it looks at the similarity of similarities – that abstracts over mechanism, feature representations and information processing. This point has been made before. For example, Haxby et al. [4] write that the disadvantage of using RSA is that:

> ...one cannot investigate whether the spaces in different subjects share the same feature tuning functions or how these tuning function codes differ for different brain regions. One cannot predict the response to a new stimulus in a subject on the basis of the responses to that stimulus in other subjects. One cannot predict the tuning function for individual neural features in terms of stimulus features, precluding investigators from predicting the response pattern vector for a new stimulus on the basis of its features. [pg. 446]

Despite these warnings, RSA continues to be used to infer that different individuals or brain regions or computational models have similar mechanism (that is, they are similar in nested functions and algorithms that transform inputs into neural response vectors). One area where these conclusions are frequently made is the comparison between the hierarchical representations in the visual cortex and Deep Neural Networks (DNNs). For example, Cichy et al. [17] observed a correspondence in the RDMs of DNNs performing object categorization and neural responses in human visual cortex recorded using MEG and fMRI. Based on this correspondence, the authors concluded that:

> ...hierarchical systems of visual representations emerge in both the human ventral and dorsal visual stream as the result of task constraints of object categorization posed in everyday life, and provide strong evidence for object

> representations in the dorsal stream independent of attention or motor intention. [pg. 5]

Thus, the correspondence in RDMs is used to infer the mechanism of emergence of visual representations. Based on a similar comparison, Kriegeskorte [19] concluded that:

> Deep convolutional feedforward networks for object recognition are not biologically detailed and rely on nonlinearities and learning algorithms that may differ from those of biological brains. Nevertheless they learn internal representations that are highly similar to representations in human and nonhuman primate IT cortex. [pg. 441]

While authors are sometimes careful in stating that the term 'similarity in representations' is used as a shorthand for a 'similarity in representational geometries', they nevertheless also invite the reader to accept that different systems show similar representational geometries because it is likely that they also use similar mechanisms to transform sensory information into latent representations, or they use similar (downstream) mechanisms to decode these latent representations. But how safe are these assumptions?

The main goal of our paper is to show that high RSA scores should not be used to infer two systems have similar mechanisms. In Study 1, in a bare-bones setup, we show that it is possible for two systems to transform input stimuli through known functions that are vastly different but end up with similar representational geometries. In particular, the study shows that 1) the presence of second-order confounds in the training data can lead systems to mimic each other's representational geometry even in the absence of mechanistic similarity, and 2) the intrinsic structure of datasets rather than mechanistic alignment can lead to artifactual modulation of RSA scores. Then in Studies 2 and 3 we show these problems extend to more complex datasets directly relevant to artificial

intelligence and computational neuroscience by making comparisons within and between sets of artificial and biological systems. Finally, in Study 4, we show that not only are misleadingly high RSA scores possible in practice but they are also highly plausible given the hierarchical structure of categories in datasets that are routinely used.

Our demonstrations provide an explanation of how these phenomena, which arise ubiquitously, can lead to incorrect inferences and contradictory or paradoxical findings. For example, it has been recently observed that correlations in representational geometries between human visual cortex and DNNs can vary from being close to the noise ceiling to being uncorrelated based on the visual stimuli used in the experiments [20]. Since our results have considerable generality with respect to current practices across multiple fields, we discuss the implications for published results, including a discussion of two alternative philosophical perspectives on the nature of mental representations that our findings speak to. We conclude by providing some general recommendations regarding how to best use RSA going forward.

Results

Proof of concept

It may be tempting to infer that two systems which have similar representational geometries for a set of concepts do so because they encode similar properties of sensory data and transform sensory data through a similar set of functions. In this section, we show that it is possible, at least in principle, for qualitatively different systems to end up with very similar representational geometries even though they (i) transform their inputs through very different functions, and (ii) select different features of inputs.

Study 1: Demonstrably different transformations of inputs can lead to low or high RSA-scores We start by considering a simple two-dimensional dataset and two systems where we know the closed-form functions that project this data into two representational spaces. This simple setup helps us gain a theoretical understanding of the circumstances under which it is possible for qualitatively different projections to show similar representational geometries.

Consider a population of animate and inanimate objects that consist of four categories of objects – birds, dogs, airplanes and bicycles. Each object in this population will have a set of stimulus features, using which one can map each exemplar from all four categories into a feature space. In Figure 2A (left), we show a hypothetical 2D feature space where exemplars from each category cluster together. Futhermore, we consider two datasets sampled from this population – Dataset A (Figure 2A, middle) which consists of birds and bicycles and Dataset B (Figure 2A, right) which consists of dogs and airplanes. Both datasets consist of animate and inanimate objects, but they differ in how items in each category are represented in the input space.

Now, consider two information-processing systems that re-represent Dataset A into two different latent spaces (Figure 2B). These could be two recognition systems designed to distinguish animate and inanimate categories. We assume that we can observe the representational geometry of the latent representations of each system and we are interested in understanding whether observing a strong correlation between these geometries implies whether the two systems have a similar *representational space* – that is, they project inputs into the latent space using similar functions. To examine this question, we consider a setup where we know the functions, Φ_1 and Φ_2, that map the inputs to the latent space in each system. We will now demonstrate that even when these functions are qualitatively different from each other, the geometry of latent representations can nev-

ertheless be highly correlated. We will also show that the difference in representational spaces becomes more clear when one considers a different dataset (Dataset B), where inputs projected using the same functions now lead to a low correlation in representational geometries.

We can compute the geometry of a set of representations by establishing the pairwise distance between all vectors in each representational space Φ. There are many different methods of computing this representational distance between any pair of vectors, all deriving from the dot product between vectors (see, for example, Figure 1 in [21]). Previous research has shown that the choice of the distance metric itself can influence the inferences one can draw from one's analysis [21, 22]. However, here our focus is not the distance metric itself, but the fundamental nature of RSA. Therefore, we use the same generic distance metric – the dot product – to compute the pair-wise distance between all vectors in both representational spaces. In other words, the representational distance $d[\Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j})]$, between the projections of any pair of input stimuli, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ into a feature space Φ, is proportional to the inner product between the projections in the feature space:

$$d[\Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j})] \;\propto\; \Phi(\boldsymbol{x_i}) \cdot \Phi(\boldsymbol{x_j}) \tag{1}$$

And we can obtain the representational geometry of the input stimuli $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ in any representational space Φ by computing the pairwise distances, $d[\Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j})]$ for all pairs of data points, (i, j). Here, we assume that the projections Φ_1 and Φ_2 are such that these pairwise distances are given by two positive semi-definite kernel functions $\kappa_1(\boldsymbol{x_i}, \boldsymbol{x_j})$ and $\kappa_2(\boldsymbol{x_i}, \boldsymbol{x_j})$, respectively:

$$d[\Phi_1(\boldsymbol{x_i}), \Phi_1(\boldsymbol{x_j})] \propto \Phi_1(\boldsymbol{x_i}) \cdot \Phi_1(\boldsymbol{x_j}) = \kappa_1(\boldsymbol{x_i}, \boldsymbol{x_j}) \quad (2)$$

$$d[\Phi_2(\boldsymbol{x_i}), \Phi_2(\boldsymbol{x_j})] \propto \Phi_2(\boldsymbol{x_i}) \cdot \Phi_2(\boldsymbol{x_j}) = \kappa_2(\boldsymbol{x_i}, \boldsymbol{x_j}) \quad (3)$$

Now, let us consider two qualitatively different kernel functions: $\kappa_1(\boldsymbol{x_i}, \boldsymbol{x_j}) = e^{\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{2\sigma^2}}$ is a radial-basis kernel (where σ^2 is the bandwidth parameter of the kernel), while $\kappa_2(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i}^T \boldsymbol{x_j}}{||\boldsymbol{x_i}||||\boldsymbol{x_j}||}$ is a cosine kernel. In other words, Φ_1 and Φ_2 are two fundamentally different projections of the inputs $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ – while Φ_2 maps a 2D input $\boldsymbol{x_i}$ into a 2D feature space, Φ_1 maps the same 2D input into an infinite-dimensional space. Nevertheless, since cosine and RBF kernels are Mercer kernels, we can compute the distances (as measured by the dot product) between each pair of projected vectors using the kernel trick [23, 24]. That is, we can find the distance between any pair of points in the representational space by applying the kernel function to those points in the input space. These pairwise distances are shown by the kernel matrices in Figure 2B.

Next, we can determine how the geometry of these projections in the two systems relate to each other by computing the correlation between the kernel matrices, shown on the right-hand-side of Figure 2B. We can see from these results that the kernel matrices are highly correlated – i.e., the input stimuli are projected to very similar geometries in the two representational spaces.

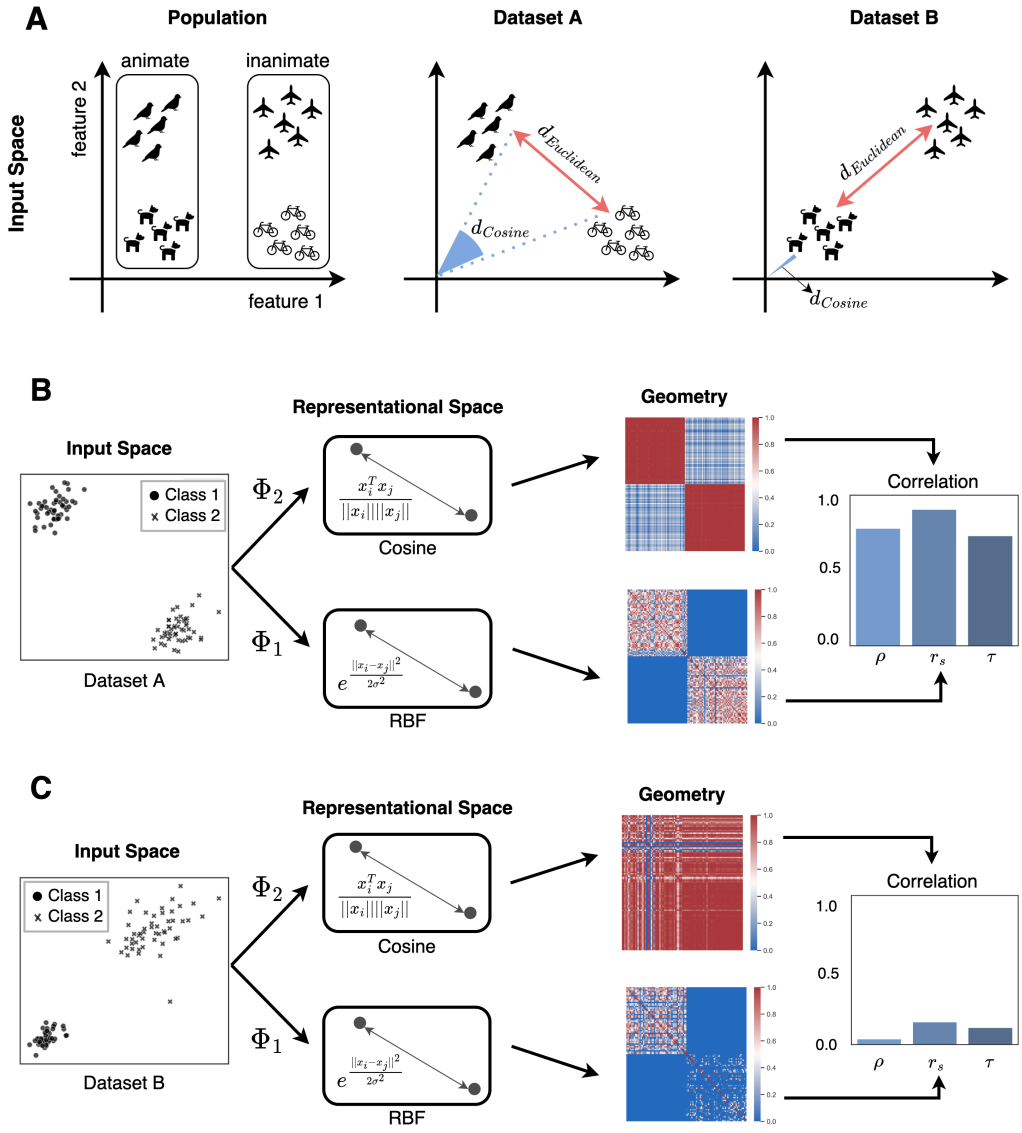

Figure 2: **Mimic and modulation effect in representational geometries.** (A) An example of a population of animate (birds, dogs) and inanimate (planes, bikes) objects, plotted in a hypothetical 2D stimulus feature space. Two datasets are sampled from this population: In Dataset A (middle), the Euclidean distance (in input space) between categories mirrors the Cosine distance, while in Dataset B (right) it does not. (B) Simulation where two systems transform stimuli in Dataset A into latent representations such that the (dot product) distance between latent vectors is given by RBF and Cosine kernels, respectively. As Euclidean and Cosine distances in the input space mirror each other, the representational geometries (visualised here using kernel matrices) end up being highly correlated (shown using Pearson (ρ), Spearman (r_s) and Kendall's (τ) correlation coefficients on the right). We call this strong correlation in representational geometries despite a difference in input transformation a *mimic effect*. (C) Simulation where objects in Dataset B are projected using same transformations as (B). The (dot product) distance is still given by the same (RBF and Cosine) kernels. However, for this dataset, the Euclidean and Cosine distances in input space do *not* mirror each other and as a consequence, the representational geometries show low correlation. Thus the correlation in representational geometries depends on how the datasets are sampled from the population. We call this change in correlation a *modulation effect*.

If one did not know the input transformations and simply observed the correlation between kernel matrices, it would be tempting to infer that the two systems Φ_1 and Φ_2 transform an unknown input stimulus $\boldsymbol{x}$ through a similar set of functions – for example functions that belong to the same class or project inputs to similar representational spaces. However, this would be an error. The projections $\Phi_1(\boldsymbol{x})$ and $\Phi_2(\boldsymbol{x})$ are fundamentally different – Φ_1 (radial basis kernel) projects an input vector into an infinite dimensional space, while Φ_2 (cosine kernel) projects it onto a unit sphere. The difference between these functions becomes apparent if one considers how this correlation changes if one considers a different set of input stimuli. For example, the set of data points from Dataset B (sampled fromt the same population) are projected to very different geometries, leading to a low correlation between the two kernel matrices (Figure 2C).

In fact, the reason for highly correlated kernel matrices in Figure 2B is not a similarity in the transformations Φ_1 and Φ_2 but the structure of the dataset. The representational distance between any two points in the first representational space, $d[\Phi_1(\boldsymbol{x_i}), \Phi_1(\boldsymbol{x_j})]$, is $e^{\frac{||\boldsymbol{x_i}-\boldsymbol{x_j}||^2}{2\sigma^2}}$. That is, the representational distance in Φ_1 is a function of their Euclidean distance $||\boldsymbol{x_i} - \boldsymbol{x_j}||$ in the input space. On the other hand, the representational distance between any two points in the second representational space, $d[\Phi_2(\boldsymbol{x_i}), \Phi_2(\boldsymbol{x_j})]$, is, $\frac{\boldsymbol{x_i}^T\boldsymbol{x_j}}{||\boldsymbol{x_i}||||\boldsymbol{x_j}||}$. That is, the representational distance in Φ_2 is a function of their cosine distance in the input space.. These two stimulus features – Euclidean distance and cosine distance – are *confounds* that lead to the same representational geometries for certain datasets. In Dataset A, the stimuli is clustered such that the Euclidean distance between any two stimuli is correlated with their cosine distance (see Figure 2A, middle). However, for Dataset B, the Euclidean distance is no longer correlated with the angle (see Figure 2A, right) and the confounds lead to different representational geometries, as can be seen in Figure 2C. Thus, this example illustrates two effects: (i) a *mimic* effect, where two sys-

tems that transform sensory input through very different functions end up with similar representational geometries (Figure 2B) , and (ii) a *modulation* effect, where two systems that are non-identical have similar representational geometries for one set of inputs, but dissimilar geometries for a second set (compare Figures 2B and 2C).

Study 2: Complex systems encoding different features of inputs can show a high RSA-score Study 1 made a number of simplifying assumptions – the dataset was two-dimensional, clustered into two categories and we intentionally chose functions Φ_1 and Φ_2 such that the kernel matrices were correlated in one case and not correlated in the other. It could be argued that, even though the above results hold in principle, they are unlikely in practice when the transformations and data structure are more complex. For example, it might be tempting to assume that accidental similarity in representational geometries becomes less likely as one increases the number of categories (i.e., clusters or conditions) being considered. However, In Figure 3 we illustrate how complex systems transforming high-dimensional input from a number of categories may achieve high RSA scores. Even though one system extracts surface reflectance and the other extracts global shape, they can end up with very similar representational geometries. This would occur if objects similar in their reflectance properties were also similar in shape (e.g., glossy balloons and light bulbs) and if objects dissimilar according to reflectance properties were also dissimilar in shape (e.g., dogs and light bulbs). This is the mimic effect, where representational geometries of these two systems end up being similar because reflectance and shape are second-order confounds in this dataset. Conducting RSA on this dataset will show a high correlation in RDMs, even though the latent representations in these systems are related to very different stimulus features.

To demonstrate this empirically, we now consider a more complex setup, where the

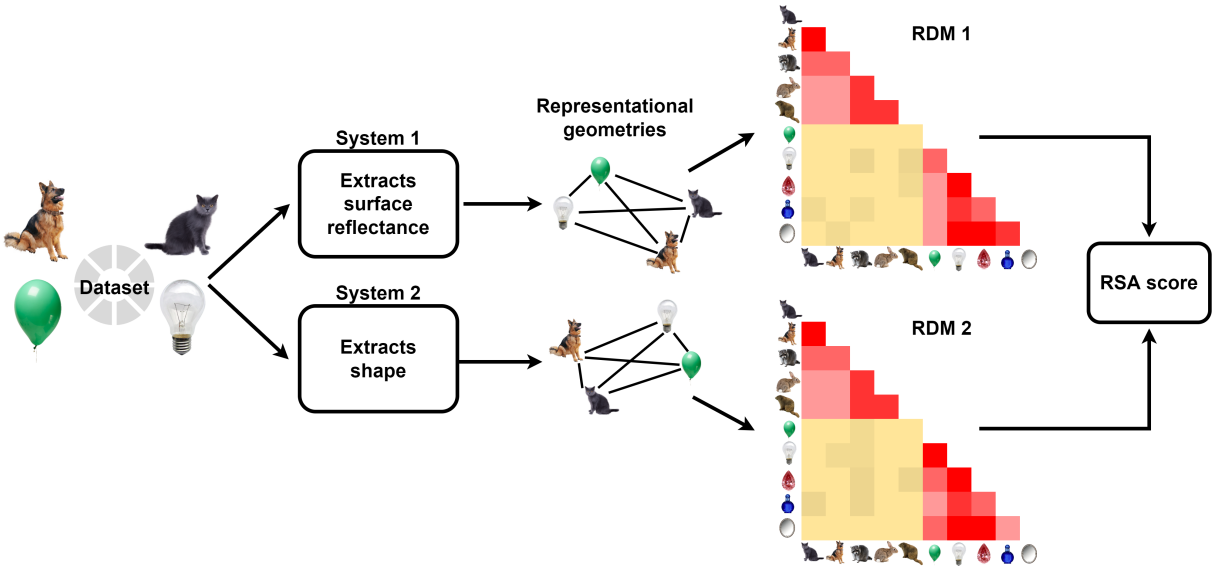

Figure 3: **Example of a second-order confound.** Two systems, one forming representations based on surface reflectance of objects (while ignoring all other features such as colour or texture) and the other based on global shape (while ignoring other features), can have very similar representational geometries. This similarity would lead to a high RSA score but would not justify an inference about the representations being similar.

transformations Φ_1 and Φ_2 are modelled as feedforward deep neural networks (DNNs), trained to classify a high-dimensional dataset into multiple categories. Many studies that use RSA compare systems using naturalistic images as visual inputs [9, 14]. While using naturalistic images brings research closer to the real-world, it is also well-known that datasets of naturalistic images frequently contain confounds – independent features that can predict image categories [25]. We will now show how the simplest of such confounds, a single pixel, can lead to a high RSA score between two DNNs that encode qualitatively different features of inputs.

Consider the same setup as above, where an input stimulus, $\boldsymbol{x}$, is transformed to a representation space by two systems, Φ_1 and Φ_2. Instead of a two-dimensional input space, $\boldsymbol{x}$ now exists in a high-dimensional image space and Φ_1 and Φ_2 are two versions of a DNN – `VGG-16` – trained to classify input images into different categories. We ensured that Φ_1 and

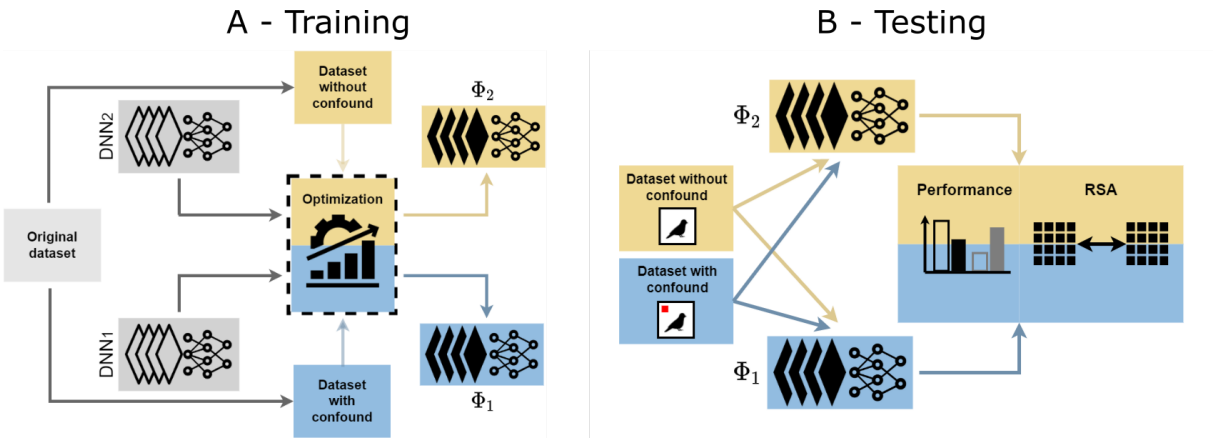

Figure 4: **Training and testing DNNs with different feature encodings.** Panel A shows the training procedure for Studies 2–4, where we created two versions of the original dataset (gray), one containing a confound (blue) and the other left unperturbed (yellow). These two datasets were used to train two networks (gray) on a categorisation task, resulting in two networks that learn to categorise images either based on the confound (projection Φ_2) or based on statistical properties of the unperturbed image (projection Φ_1). Panel B shows the testing procedure where each network was tested on stimuli from each dataset – leading to a 2x2 design. Performance on these datasets was used to infer the features that each network encoded and their internal response patterns were used to calculate RSA-scores between the two networks.

Φ_2 were qualitatively different transformations of input stimuli by making the networks sensitive to different predictive features within the stimuli. The first network was trained on an unperturbed dataset, while the second network was trained on a modified version of the dataset, where each image was modified to contain a confound – a single pixel in a location that was diagnostic of the category (see Figure 4 for the general approach).

The locations of these diagnostic pixels were chosen such that they were correlated to the corresponding representational distances between classes in Φ_1. Our hypothesis was that if the representational distances in Φ_2 preserve the physical distances of diagnostic pixels in input space, then this confound will end up mimicking the representational geometry of Φ_1, even though the two systems use qualitatively different features for classification. Furthermore, we trained two more networks, Φ_3 and Φ_4, which were identical

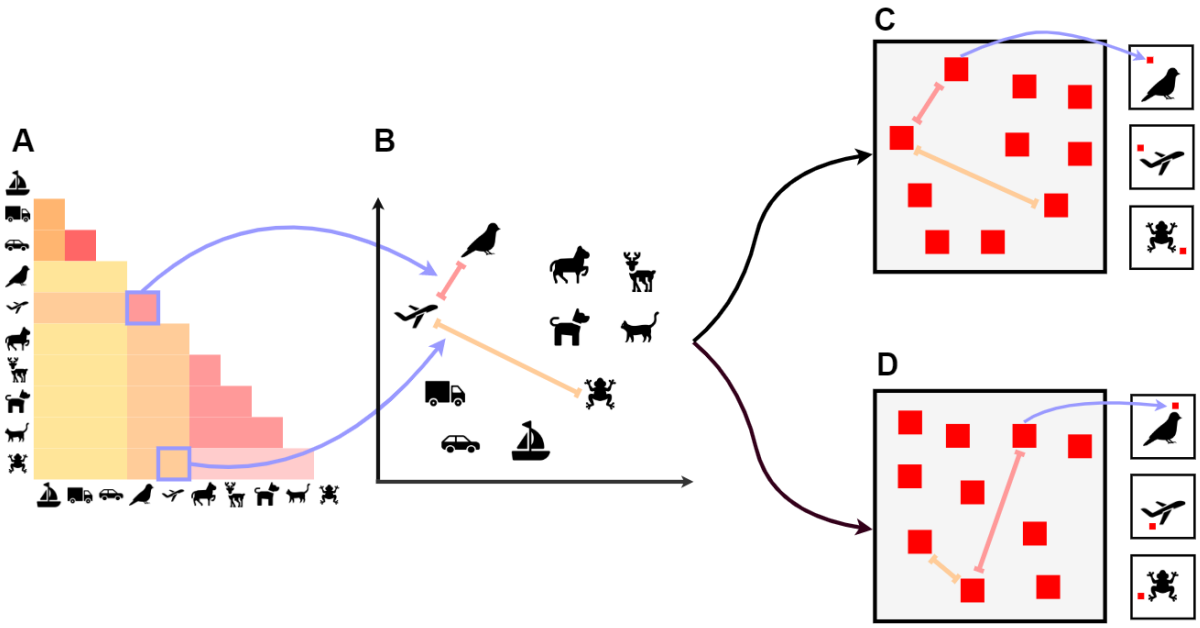

Figure 5: **Study 2 confound placement.** The representational geometry (Panel A and B) from the network trained on the unperturbed `CIFAR-10` images is used to determine the location of the single pixel confound (shown as a red patch here) for each category. In the 'Positive' condition (Panel C), we determined 10 locations in a 2D plane such that the distances between these locations were positively correlated to the representational geometry – illustrated here as the red patches in Panel C being in similar locations to category locations in Panel B. These 10 locations were then used to insert a single diagnostic – i.e., category-dependent – pixel in each image (Insets in Panel C). A similar procedure was also used to generate datasets where the confound was uncorrelated (Panel D) or negatively correlated (not shown here) with the representational geometry of the network.

to Φ_2, except these networks were trained on datasets where the location of the confound was uncorrelated (Φ_3) or negatively correlated (Φ_4) with the representational distances in Φ_1 (see Figure 5 and Methods for details).

Classification accuracy (Figure 6 (left)) revealed that the network Φ_1, trained on the unperturbed images, learned to classify these images and ignored the diagnostic pixel – that is, it's performance was identical for the unperturbed and modified images. In contrast, networks Φ_2 (positive), Φ_3 (uncorrelated) and Φ_4(negative) failed to classify the unperturbed images (performance was near chance) but learned to perfectly classify the modified images, showing that these networks develop qualitatively different representations compared to normally trained networks.

Next we computed pairwise RSA scores between the representations at the last convolution layer of Φ_1 and each of Φ_2, Φ_3 and Φ_4 (Figure 6 (right)). When presented unperturbed test images, the Φ_2, Φ_3 and Φ_4 networks all showed low RSA scores with the normally trained Φ_1 network. However, when networks were presented with test images that included the predictive pixels, RSA varied depending on the geometry of pixel locations in the input space. When the geometry of pixel locations was positively correlated to the normally trained network, RSA scores approached ceiling (i.e., comparable to RSA scores between two normally trained networks). Networks trained on uncorrelated and negatively correlated pixel placements scored much lower.

These results mirror Study 1: we observed that it is possible for two networks (Φ_1 and Φ_2) to show highly correlated representational geometries even though these networks learn to classify images based on very different features. One may argue that this could be because the two networks could have learned similar representations at the final convolution layer of the DNN and it is the classifier that sits on top of this representation that leads to the behavioural differences between these networks. But if this was true, it

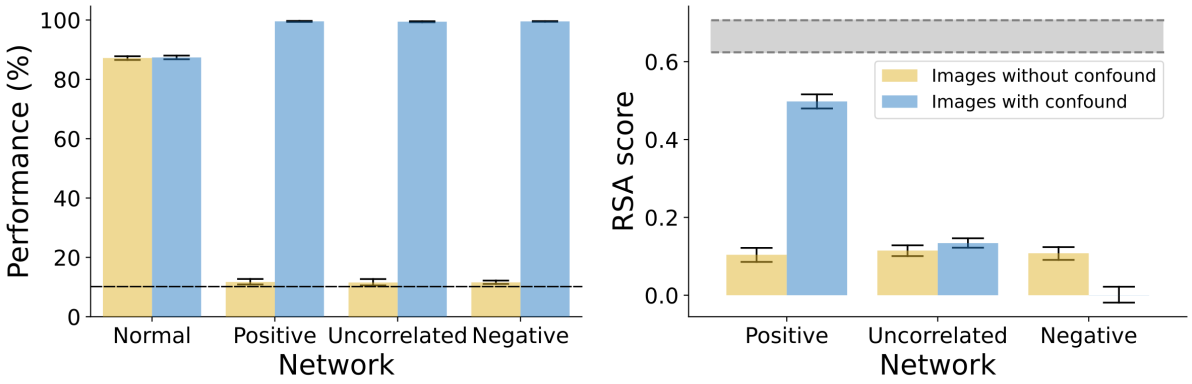

Figure 6: **Study 2 results.** *Left:* Performance of normally trained networks did not depend on whether classification was done on unperturbed `CIFAR-10` images or images with a single pixel confound (error bars represent 95% CI, the dashed line represents chance performance). All three networks trained on datasets with confounds could perfectly categorise the test images when they contained the confound (blue bars), but failed to achieve above-chance performance if the predictive pixel was not present (yellow bars). *Right:* The RSA score between the network trained on the unperturbed dataset and each of the networks trained on datasets with confounds. The three networks showed similar scores when tested on images without confounds, but vastly different RSA scores when tested on images with confounds. Networks in the Positive condition showed near ceiling scores (the shaded area represents noise ceiling) while networks in the Uncorrelated and Negative conditions showed much lower RSA.

would not explain why RSA scores diminish for the two other comparisons (with Φ_3 and Φ_4). This modulation of RSA-scores for different datasets suggests that, like in Study 1, the correlation in representational geometry is not because the two systems encode similar features of inputs, but because different features mimic each other in their representational geometries.

Re-examining some influential findings

In Studies 1 and 2, we showed that it is possible for qualitatively different systems to end up with similar representational geometries. However, it may be argued that while this is possible in principle, it is unlikely in practice in real-world scenarios. In the following two studies, we consider real-world data from some recent influential experiments, recorded from both primate and human cortex. We show how RSA-scores can be driven by confounds in these real-world settings and how properties of training and test data may contribute to observed RSA-scores.

Study 3: Neural activations in monkey IT cortex can show a high RSA-score with DNNs despite different encoding of input data In our next study, we consider data from experiments comparing representational geometries between computational models and macaque visual cortex [14, 26]. The experimental setup was similar to Study 2, though note that unlike Study 2, where both systems used the same architecture and learning algorithm, this study considered two very different systems – one artificial (DNN) and the other biological (macaque IT cortex). We used the same set of images that were shown to macaques by Majaj et al. [27] and modified this dataset to superimpose a small diagnostic patch on each image. In the same manner as in Study 2 above, we constructed three different datasets, where the locations of these diagnostic

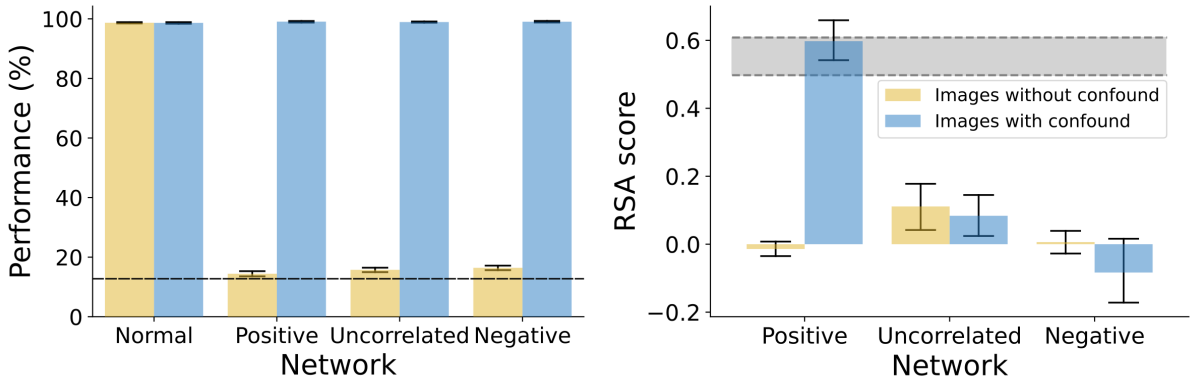

Figure 7: **Study 3 results.** *Left:* Classification Performance of the network trained on unperturbed images (Normal condition) did not depend on the presence or absence of the confound, while performance of networks trained with the confound (Positive, Uncorrelated and Negative conditions) highly depended on whether the confound was present (dashed line represents chance performance). *Right:* RSA-scores with macaque IT activations were low for all three conditions when images did not contain a confound (yellow bars). When images contained a confound (blue bars), the RSA-scores depended on the condition, matching the RSA-score of the normally trained network (grey band) in the Positive condition, but decreasing significantly in the Uncorrelated and Negative conditions. The grey band represents a 95% CI for the RSA-score between normally trained networks and macaque IT activations.

patches were either positively correlated, uncorrelated or negatively correlated with the RDM of macaque activations. We then trained four CNNs. The first CNN was pre-trained on `ImageNet` and then fine-tuned on the unmodified dataset of images shown to the macaques. Previous research has shown that CNNs trained in this manner develop representations that mirror the representational geometry of neurons in primate inferior temporal (IT) cortex [14]. The other three networks were trained on the three modified datasets and learned to entirely rely on the diagnostic patches (accuracy on images without the diagnostic patches was around chance).

Figure 7 (right) shows the correlation in representational geometry between the macaque IT activations and activations at the final convolution layer for each of these networks. The correlation with networks trained on the unmodified images is our baseline and shown as the gray band in Figure 7. Our first observation was that a CNN trained to rely on

the diagnostic patch can indeed achieve a high RSA score with macaque IT activations. In fact, the networks trained on patch locations that were positively correlated to the macaque RDM matched the RSA score of the CNNs trained on `ImageNet` and the unmodified dataset. This shows how two systems having very different architectures, encoding fundamentally different features of inputs (single patch vs naturalistic features) can show a high correspondence in their representational geometries. We also observed that, like in Study 2, the RSA score depended on the clustering of data in the input space – when patches were placed in other locations (uncorrelated or negatively correlated to macaque RDMs) the RSA score became significantly lower.

Study 4: High RSA-scores may be driven by the structure of testing data All the studies so far have used the same method to construct datasets with confounds – we established the representational geometry of one system (Φ_1) and constructed datasets where the clustering of features (pixels) mirrored this geometry. However, it could be argued that confounds which cluster in this manner are unlikely in practice. For example, even if texture and shape exist as confounds in a dataset, the inter-category distances between textures are not necessarily similar to the inter-category distances between shape.

However, categories in real-world datasets are usually hierarchically clustered into higher-level and lower-level categories. For example, in the `CIFAR-10` dataset, the Dogs and Cats (lower-level categories) are both animate (members of a common higher-level category) and Airplanes and Ships (lower-level categories) are both inanimate (members of a higher-level category). Due to this hierarchical structure, Dog and Cat images are likely to be closer to each other not only in their shape, but also their colour and texture (amongst other features) than they are to Airplane and Ship images. In our next simulation, we explore whether this hierarchical structure of categories can lead to a correlation

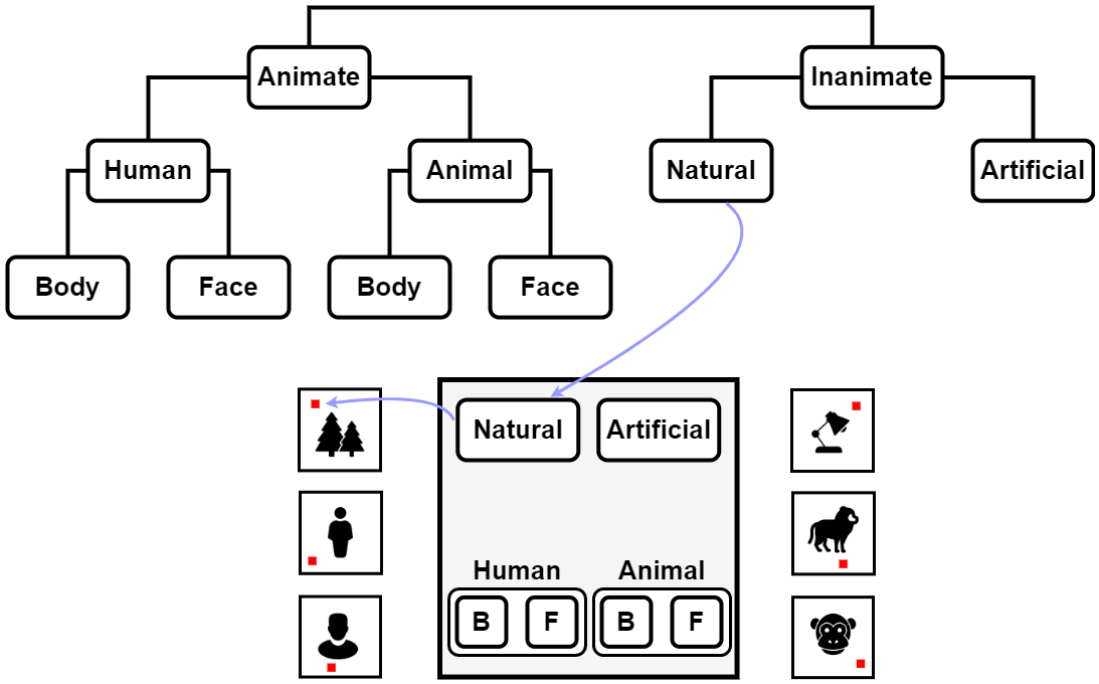

Figure 8: **Exploiting intrinsic dataset hierarchy in order to place confounds.** The top panel shows the hierarchical structure of categories in the dataset, which was used to place the single pixel confounds. The example at the bottom (middle) shows one such hierarchical placement scheme where the pixels for Inanimate images were closer to the top of the canvas while Animate images were closer to the bottom. Within the Animate images, the pixels for Humans and Animals were placed at the left and right, respectively, and the pixels for bodies (B) and faces (F) were clustered as shown.

in representational geometries between two systems that learn different feature encodings.

For this study, we selected a popular dataset used for comparing representational geometries in humans, macaques and deep learning models [15, 28]. This dataset consists of six categories which can be organised into a hierarchical structure shown in Figure 8. [9] showed a striking match in RDMs for response patterns elicited by these stimuli in human and macaque IT. For both humans and macaques, distances in response patterns were larger between the higher-level categories (animate and inanimate) than between the lower-level categories (e.g., between human bodies and human faces).

We used a similar experimental paradigm to the above studies, where we trained

networks to classify stimuli which included a single predictive pixel. But instead of using an RDM to compute the location of a diagnostic pixel, we used the hierarchical categorical structure. In the first modified version of the dataset, the location of the pixel was based on the hierarchical structure of categories in Figure 8 – predictive pixels for animate kinds were closer to each other than to inanimate kinds, and pixels for faces were closer to each other than to bodies, etc. One such configuration can be seen in Figure 8. In the second version, the predictive pixel was placed at a random location for each category (but, of course, at the same location for all images within each category). We call these conditions 'Hierarchical' and 'Random'. [15] showed that the RDM of average response patterns elicited in the human IT cortex (Φ_1) correlated with the RDM of a DNN trained on naturalistic images (Φ_2). We explored how this compared to the correlation with the RDM of a network trained on the Hierarchical pixel placement (Φ_3) and Random pixel placement (Φ_4).

Results for this study are shown in Figure 9. We observed that representational geometry of a network trained on Hierarchically placed pixels (Φ_3) was just as correlated to the representational geometry of human IT responses (Φ_1) as a network trained on naturalistic images (Φ_2). However, when the pixel locations for each category were randomly chosen, this correlation decreased significantly. These results suggest that any confound in the dataset (including texture, colour or low-level visual information) that has distances governed by the hierarchical clustering structure of the data could underlie the observed similarity in representational geometries between CNNs and human IT. More generally, these results show how it is plausible that many confounds present in popular datasets may underlie the observed similarity in representational geometries between two systems. The error of inferring a similarity in mechanism based on a high RSA score is not just possible but also probable.

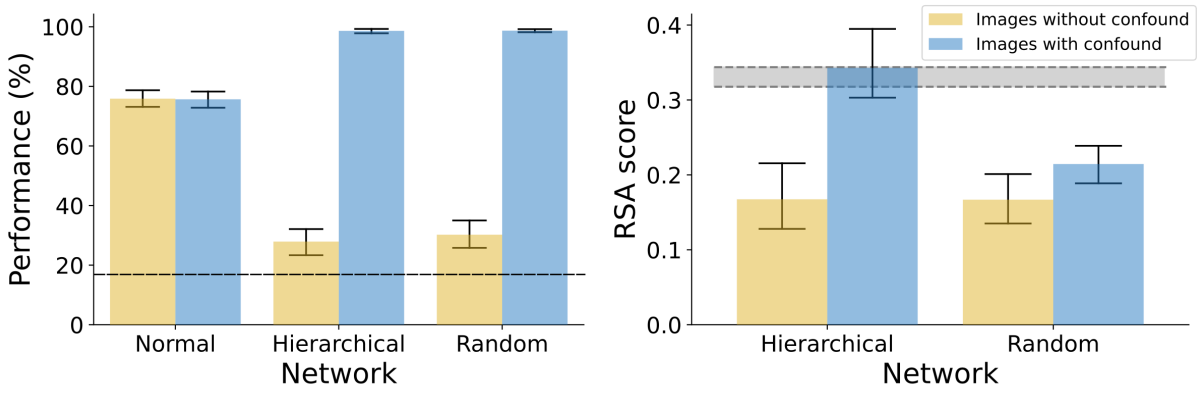

Figure 9: **Study 4 results.** *Left:* Performance of normally trained networks did not depend on whether the confound was present. Networks trained with the confound failed to classify stimuli without the confound (yellow bars) while achieving near perfect classification of stimuli with the confound present (blue bars, dashed line represents chance performance). *Right:* RSA with human IT activations reveals that, when the confound was present, the RSA-score for networks in the Hierarchical condition matched the RSA-score of normally trained network (gray band), while the RSA-score of the network in the Random condition was significantly lower. The grey band represents 95% CI for the RSA score between normally trained networks and human IT.

Discussion

In four studies, we have illustrated a number of conditions under which it can be problematic to infer a similarity of representations between two systems based on a correlation in their representational geometries. In particular, we showed that two systems may transform their inputs through very different functions and encode very different features of inputs and yet have highly correlated representational geometries. Of course, one may acknowledge that a second-order isomorphism of activity patterns does *not* strictly imply that two systems are similar mechanistically but still assume that it is highly likely to be the case. That is, as a practical matter, a researcher may assume that RSA is a reliable method to compare systems. However, our findings challenge this assumption. We show how a high RSA score between different systems can not only occur in a bare-bones simulation (Study 1), but also in practice, in high-dimensional systems operating on high-

dimensional data (Studies 2–3). Furthermore, we show that the hierarchical structure of datasets frequently used to test similarity of representations lends itself to a high RSA score arising because of second-order confounds present in the dataset (Study 4). Therefore, second-order confounds driving high RSA scores is not only possible but plausible.

One limitation of our method is that we manually insert a confound in input stimuli (in Studies 2–4) and train a network based on this confound. Even though our findings demonstrate that second-order confounds are plausible, they do not allow us to infer whether such confounds *are* present in existing datasets and driving the observed similarity in existing studies. In our view, there are two methods one could use to check whether confounds are driving results of RSA. The best way would be to identify the stimulus features in a dataset that mimic each other in representational space (e.g. shape and reflectance in Figure 3). This is not straightforward to do in high-dimensional stimuli, such as naturalistic images, which consist of millions of features. However, another approach is more tractable: conduct controlled experiments to establish whether the two systems are representing information in similar ways. We have argued for this approach in relation to making inferences about mechanistic similarity between DNNs and humans [29]. In fact, research relating DNNs to human vision provides a striking case of a disconnect between RSA and behavioural findings from psychology [29–31]. The findings here may explain contradictory RSA scores between DNNs and human visual processing as pointed out by Xu and Vaziri-Pashkam [20]. At the very least, a researcher claiming that two systems are mechanistically similar to one another based on high RSA scores should have an explanation for this discrepancy.

A related point has been made by Kriegeskorte and Diedrichson [32] and Kriegeskorte and Wei [33], who point out that two systems may have the same representational geometry, even if they have a different activity profile over neurons. In this sense, the geometry

abstracts away the information about how information was distributed over a set of neurons. Kriegeskorte and Diedrichson [32] equate this loss in information to "peeling a layer of an onion" – downstream decoders that are sensitive to the representational geometry rather than activity profiles over neuron populations can focus on difference in information as reflected by a change in geometry and be agnostic to how this information is distributed over a set of neurons. We agree that this invariance over activity profiles is indeed a useful property of representational geometries for downstream decoders. However, we are not aware of any studies that highlight how representational geometries also abstract over behaviourally relevant stimulus properties . While abstracting over activity profiles may be useful, abstracting over stimulus properties loses an important piece of information when comparing representations across brain regions, individuals, species and between brains and computational models. Our studies show how two systems may appear similar based on their representational geometries in one circumstance (e.g. Figure 2B) but drastically different in another circumstance (Figure 2C).

It is important to note how our results differ from previous studies exploring limitations of RSA. A number of studies have focused on the importance of how neural data is pre-processed and how the distance between neural patterns is computed. For example, Ramirez [34] found that pre-processing steps, such as centering (de-meaning) activation vectors may lead to incorrect inference about the representational geometry of activations. He demonstrated that subtracting the mean from activations could change the rank order of similarity between conditions. In turn, this could lead to clearly distinct RDMs becoming highly correlated and vice-versa. While this is an important methodological point, it is clearly distinct from the point we are making in this study. Indeed, the results here are agnostic of the data pre-processing steps and hold whether or not activations are centered.

Some previous studies have also explored how confounds present in data can influence the results of RSA. For example, Henriksson et al. [35] and Cai et al. [36] demonstrated that RDMs measured based on fMRI data can be severely biased because of temporal and spatial correlations in neural activity. These authors have pointed out that if activity patterns from different brain regions are recorded during the same trial, the similarity estimates will be exaggerated due to correlated neural fluctuations in these regions. Similarly, neural activity is correlated over time, which means estimated similarity based on activity patterns from the same imaging run also introduces a strong bias in RDMs. These sources of bias are important to understand, but they can also be addressed by a more careful task design and analysis [36]. In contrast, the confounds that are highlighted in this study exist in the stimulus itself. Therefore, even if one were to completely mitigate the bias in estimating RDMs, the types of confounds we highlight in our work would still pose problems when drawing inferences from correlation in RDMs.

Similarly, previous research has also highlighted the importance of choosing the correct distance metric when using RSA. For example, Ramirez [22] compared Euclidean distance with an angular metric (such as cosine similarity) and showed that the choice of distance metric can reveal different aspects of the same fMRI data. They argued that the Euclidean distance is particularly sensitive to the mean activity over a recorded voxel. Based on this analysis, Ramirez [22] suggested using an angular distance metric, especially when neural signal is aggregated over large number of neurons. Similarly, in another exhaustive study over distance measures, Bobadilla-Suarez et al. [21], evaluated neural similarity using various distance measures, including angle-based measures (cosine, Pearson, Spearman) and magnitude-based measures (Euclidean, Mahalanobis, Minkowski) and found that the choice of metric significantly influenced the measured similarity. They also found that there was no one metric that outperformed all others – rather, the preferred metric varied

across different studies, but was consistent across brain regions within a study. The choice of distance metric is again a related but orthogonal issue to the one we highlight in this study. Our results show that representational geometry loses information about stimulus features and different stimulus features (and indeed transformations of input stimulus) can lead to the same geometry. This is fundamental to the nature of representational geometries, rather than a consequence of the distance metric used.

Of course, the problem of confounds in stimuli is not unique to RSA and will affect other statistical analyses, including multivariate regression methods such as MVP classification. Indeed, the problem of confounds in stimuli is well appreciated in many different contexts [25, 37, 38], but there has been no consideration of whether these confounds are contributing to RSA findings. Perhaps this is because, unlike for MVP classification, a confound for RSA needs to not only help decode category membership, but also lead to a second-order isomorphism. Nevertheless, as we illustrate in Figure 3, there could be confounds with a second-order similarity structure in many datasets that are the product of unexpected properties of the world or the product of how these datasets are curated or hierarchically organized. This is problematic as we have clearly shown that these second order confounds can drive high RSA scores.

A reader could ask why these results matter. Couldn't a researcher take the view that representational geometry *is* representation and therefore, a strong correlation in representational geometries between two systems is sufficient to infer that the systems are representing the world in a similar manner? This question goes to the heart of an existing debate in philosophy, where philosophers distinguish between the *externalist* and *holistic* views on mental representations. According to the first view, the content of representations is determined by their relationship to entities in the external world. This perspective is implicitly taken by most neuroscientists and psychologists, who are interested in com-

paring mechanisms underlying cognitive processes – that is, they are interested in the set of nested functions and algorithms responsible for transforming sensory input into a set of activations in the brain. From this perspective, our finding that high RSA scores can be obtained between systems that work in qualitatively different ways poses a challenge to researchers using RSA to compare systems.

Alternatively, a researcher may reject an externalist view and adopt the perspective that representations obtain their meaning based on how they are related to each other within each system, rather than based on their relationship to entities in the external world. That is, "representation *is* the representation of similarities" [39]. From this perspective, as long as the two systems share the same relational distances between internal activations, one can validly infer that the two systems have similar representations. That is, a second-order isomorphism implies a similarity of representations, by definition. This view has been called *holism* in the philosophy of mind [40, 41] and is related to a similar idea of *meaning holism* in language, which is the idea that the meaning of a linguistic expression is determined by its relation to other expressions within a language [42, 43]. For example, Firth [44] (p. 11) writes: "you shall know a word by the company it keeps". Similarly, Griffiths and Steyvers [45], and Griffiths, Steyvers, and Tenenbaum [46] have adopted meaning holism accounts of semantic representations in neural networks. More recently, Piantadosi and Hill [47] have argued that large language models capture important aspects of meaning and approximate human cognition because they represent relations between concepts and their roles within a representational geometry. Even if a researcher was to adopt this holistic perspective on representations, our results should still be of interest to them as they show that the similarity between representational geometries can vary based on the visual stimulus that is used to compare them (the modulation effect). Additionally, our results show that adopting this view misses the information about

differences in mechanistic processes that a psychologist or neuroscientist is frequently interested in, for instance, whether the visual system processes surface reflectance or shape (or the location of diagnostic pixels) in order to identify objects. Fodor and Lepore long ago criticized this philosophical stance [41,48], and interestingly, this philosophical debate played an important part in the development of RSA (see Supplementary Information, Section A). Unfortunately, this debate has largely been ignored by researchers who use RSA as a method to compare similarity of systems.

In closing, we describe our recommendations for practitioners who find RSA to be useful for their research goals. These will be especially relevant to the large majority of researchers in computational, cognitive, and systems neuroscience, cognitive scientists and AI practitioners, who are interested in mechanistic similarities (i.e., they adopt an externalist position). But they should also be relevant to adopters of the holistic view who are interested in how observed representational geometries depend on the stimulus used to extract them.

First, since the intrinsic structure of datasets can artificially modulate RSA scores, researchers should compare systems on a wider variety of datasets and sampling schemes than currently done. Second, given that confounding features can lead to mimicked representational geometries, researchers should consider running additional controlled experiments to rule out this possibility when inferences hinge crucially on it. Third, when studies are conducted to search for evidence of mechanistic similarity between two or more systems, researchers should use a wider range of complementary methods, each addressing the others' blindspots (e.g., RSA combined with neural predictivity [14], MVPC [6, 49], CCA [50], SVCCA [51], CKA [52]).

Lastly, perhaps the most important general recommendation we make is that researchers should acknowledge, procedurally and in writing, which inferences are afforded

by the use of RSA, and what dissimilarities remain possible despite having observed a given pattern of RSA scores. To this end, we believe that general statements of similarity tend to obfuscate rather than accurately summarize any set of RSA-based results. Instead, we urge researchers using RSA (1) to justify the use of this method by theoretically motivated interest in representational geometry or otherwise consider other tools that best fit their goals, and (2) to state in precise terms that RSA scores reflect the similarity of representational geometries in particular, and generally avoid underspecified claims of similarity.

Methods

Dataset generation and training

All DNN simulations (Studies 2–4) were carried out using the `Pytorch` framework [53]. The model implementations were downloaded from the `torchvision` library. Networks trained on unperturbed datasets in all studies were pre-trained on `ImageNet` as were networks trained on modified datasets in Study 2. Networks trained on modified datasets in Studies 3 and 4 were randomly initialised. For the pre-trained models, their pre-trained weights were downloaded from `torchvision.models` subpackage.

Study 1 Each dataset in Study 1 consists of 100 samples (50 in each cluster) drawn from two multivariate Gaussians, $\mathcal{N}(x|\mu, \mathbf{\Sigma})$, where μ is a 2-dimensional vector and $\mathbf{\Sigma}$ is a 2×2 covariance matrix. In Figure 2A, the two Gaussians have means $\mu_\mathbf{1} = (1, 8)$ and $\mu_\mathbf{2} = (8, 1)$ and a covariance matrices $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \frac{1}{2}\mathbf{I}$, while in Figure 2B the Gaussians have means $\mu_\mathbf{1} = (1, 1)$ and $\mu_\mathbf{2} = (8, 8)$ and a covariance matrices $\mathbf{\Sigma_1} = \mathbf{I}$, $\mathbf{\Sigma_2} = 8\mathbf{I}$. All kernel matrices were computed using the `sklearn.metrics.pairwise` module of the

`scikit-learn` Python package.

Study 2 First, a VGG-16 deep convolutional neural network [54], pre-trained on the `ImageNet` dataset of naturalistic images, was trained to classify stimuli from the `CIFAR-10` dataset [55]. The `CIFAR-10` dataset includes 10 categories with 5000 training, and 1000 test images per category. The network was fine-tuned on `CIFAR-10` by replacing the classifier so that the final fully-connected layer reflected the correct number of target classes in `CIFAR-10` (10 for `CIFAR-10` as opposed to 1000 for `ImageNet`). Images were rescaled to a size of 224×224px and then the model learnt to minimise the cross-entropy error using the RMSprop optimizer with a mini-batch size of 64, learning rate of 10^{-5}, and momentum of 0.9. All models were trained for 10 epochs, which were sufficient for convergence across all datasets.

Second, 100 random images from the test set for each category were sampled as input for the network and activations at the final convolutional layer extracted using the `THINGSVision` Python toolkit [56]. The same toolkit was used to generate a representational dissimilarity matrix (RDM) from the pattern of activations using `1-Pearson's r` as the distance metric. The RDM was then averaged by calculating the median distance between each instance of one category with each instance of the others (e.g., the median distance between `Airplane` and `Ship` was the median of all pair-wise distances between activity patterns for airplane and ship stimuli). This resulted in a 10×10, category-level, RDM which reflected median between-category distances.

Third, three modified versions of the `CIFAR-10` datasets were created for the 'Positive', 'Uncorrelated' and 'Negative' conditions, respectively. In each dataset, we added one diagnostic pixel to each image, where the location of the pixel depended on the category (See Figure 5). The locations of these pixels were determined using the averaged RDM

from the previous step. We call this the target RDM. In the 'Positive' condition, we wanted the distances between pixel placements to be positively correlated to the distances between categories in the target RDM. We achieved this by using an iterative algorithm that sampled pixel placements at random, calculated an RDM based on distances between the pixel placements and computed an RSA score (Spearman correlation) with the target RDM. Placements with a score above 0.70 were retained and further optimized (using small perturbations) to achieve an RSA-score over 0.90. The same procedure was also used to determine placements in the Uncorrelated (optimizing for a score close to 0) and Negatively correlated (optimizing for a negative score) conditions.

Finally, datasets were created using 10 different placements in each of the three conditions. Networks were trained for classification on these modified `CIFAR-10` datasets in the same manner as the VGG-16 network trained on the unperturbed version of the dataset (See Figure 4).

Study 3 The procedure mirrored Study 2 with the main difference being that the target system was the macaque inferior temporal cortex. Neural data from two macaques, as well as the dataset were obtained from the Brain Score repository [26]. This dataset consists of 3200 images from 8 categories (animals, boats, cars, chairs, faces, fruits, planes, and tables), we computed an 8×8 averaged RDM based on macaque IT response patterns for stimuli in each category.

This averaged RDM was then used as the target RDM in the optimization procedure to determine locations of the confound (here, a white predictive patch of size 5×5 pixels) for each category. Using a patch instead of a single pixel was required in this dataset because of the structure and smaller size of the dataset (3200 images, rather than 50,000 images for `CIFAR-10`). In this smaller dataset, the networks struggle to learn based on a single

pixel. However, increasing the size of the patch makes these patches more predictive and the networks are able to again learn entirely based on this confound (see results in Figure 6). In a manner similar to Study 2, this optimisation procedure was used to construct three datasets, where the confound's placement was positively correlated, uncorrelated or negatively correlated with the category distances in the target RDM.

Finally, each dataset was split into 75% training (2432 images) and 25% test sets (768 images) before VGG-16 networks were trained on the unperturbed and modified datasets in the same manner as in Study 2. One difference between Studies 2 and 3 was that here the networks in the Positive, Uncorrelated and Negative conditions were trained from scratch, i.e., not pre-trained on `ImageNet`. This was done because we wanted to make sure that the network in the Normal condition (trained on `ImageNet`) and the networks in the Positive, Uncorrelated and Negative conditions encoded fundamentally different features of their inputs – i.e., there were no `ImageNet`-related features encoded by representations Φ_2, Φ_3 and Φ_4 that were responsible for the similarity in representational geometries between these representations and the representations in macaque IT cortex.

Study 4 The target system in this study was human IT cortex. The human RDM and dataset were obtained from [9]. Rather than calculating pixel placements based on the human RDM, the hierarchical structure of the dataset was used to place the pixels manually. The dataset consists of 910 images from 6 categories: human bodies, human faces, animal bodies, animal faces, artificial inanimate objects and natural inanimate objects. These low-level categories can be organised into the hierarchical structure shown in Figure 8. Predictive pixels were manually placed so that the distance between pixels for Animate kinds were closer together than they were to Inanimate kinds and that faces were closer together than bodies. This can be done in many different ways, so we created

five different datasets, with five possible arrangements of predictive pixels. Results in the Hiearchical condition (Figure 9) are averaged over these five datasets. Placements for the Random condition were done similarly, except that the locations were selected randomly.

Networks were then trained on a 6-way classification task (818 training images and 92 test images) in a similar manner to the previous studies. As in Study 3, networks trained on the modified datasets (both Hierarchical and Random conditions) were not pre-trained on `ImageNet`.

RDM and RSA computation

For Studies 2-4 all image-level RDMs were calculated using $1-r$ as the distance measure. RSA scores were computed as the Spearman rank correlation between RDMs.

In Study 2, a curated set of test images was selected due to the extreme heterogeneity of the `CIFAR-10` dataset (low activation pattern similarity between instances of the same category). This was done by selecting 5 images per category which maximally correlated with the averaged activation pattern for the category. Since `CIFAR-10` consists of 10 categories, the RSA-scores in Study 2 were computed using RDMs of size 50×50.

In Study 3, the dataset consisted of 3200 images belonging to 8 categories. We first calculated a full 3200×3200 RDM using the entire set of stimuli. An averaged, category-level, 8×8 RDM was then calculated using median distances between categories (in a manner similar to that described for Study 2 in the Section 'Dataset generation and training'). This 8×8 RDM was used to determine the RSA-scores. We also obtained qualitatively similar results using the full 3200×3200 RDMs. These results can be found in the Supplementary Information, Section B.

In Study 4, the dataset consisted of 818 training images and 92 test images. Kriegesko-

rte et al. [9] used these images to obtain a 92×92 RDM to compare representations between human and macaque IT cortex. Here we computed a similar 92 × 92 RDM for networks trained in the Normal, Hierarchical and Random training conditions, which were then compared with the 92 × 92 RDM from human IT cortex to obtain RSA-scores for each condition.

Testing

In Study 2, we used a 4 × 2 design to measure classification performance for networks in all four conditions (Normal, Postive, Uncorrelated and Negative) on both unperturbed images and modified images. We computed six RSA-scores: three pairs of networks – Normal-Positive, Normal-Uncorrelated and Normal-Negative – and two types of inputs – unperturbed and modified test images. The noise ceiling (grey band in Figure 6) was determined in the standard way as described in [57] and represents the expected range of the highest possible RSA score with the target system (network trained on the unperturbed dataset).

In Study 3, performance was estimated in the same manner as in Study 2 (using a 4×2 design), but RSA-scores were computed between RDMs from macaque IT activations and the four types of networks – i.e. for the pairs Macaque-Normal, Macaque-Positive, Macaque-Uncorrelated and Macaque-Negative. And like in Study 2, we determined each of these RSA-scores for both unperturbed and modified test images as inputs to the networks.

In Study 4, performance and RSA were computed in the same manner as in Studyn 3, except that the target RDM for RSA computation came from activations in human IT cortex and the networks were trained in one of three conditions: Normal, Hierarchical and Random.

Data analysis

Performance and RSA scores were compared by running analyses of variance and Tukey HSD post-hoc tests. In Study 2 and 3, performance differences were tested by running a 4 (type of training) by 2 (type of dataset) mixed ANOVAs. In, Study 4, the differences were tested by running a 3x2 mixed ANOVA.

RSA scores with the target system between networks in various conditions were compared by running 3x2 ANOVAs in Studies 2 and 3, and a 2x2 ANOVA in Study 4. We observed that RSA-scores were highly dependent on both the way the networks were trained and also the test images used to elicit response activations.

For a detailed overview of the statistical analyses and results, see Supplemental Information Section C.

Data Availability

Confound placement coordinates (Studies 2-4), unperturbed datasets (Studies 3 and 4), macaque activation patterns and RDMs (Study 3) and human RDM (Study 4) are available at OSF.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 741134)

References

[1] Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology* **148**, 574–591 (1959).

[2] O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Experimental Neurology* **51**, 78–109 (1976).

[3] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).

[4] Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* **37**, 435–456 (2014).

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

[6] Haxby, J. *et al.* A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011). URL https://www.sciencedirect.com/science/article/pii/S0896627311007811.

[7] Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17**, 401–412 (2013). URL https://www.sciencedirect.com/science/article/pii/S1364661313001277.

[8] Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2** (2008).

[9] Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).

[10] O'Hearn, K., Larsen, B., Fedor, J., Luna, B. & Lynn, A. Representational similarity analysis reveals atypical age-related changes in brain regions supporting face and car recognition in autism. *NeuroImage* **209**, 116322 (2020).

[11] Michael L. Mack, B. L., Alison R. Preston. Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology* **23**, 2023–2027 (2013).

[12] Freund, M. C., Etzel, J. A. & Braver, T. S. Neural coding of cognitive control: The representational similarity analysis approach. *Trends in Cognitive Sciences* **25**, 622–638 (2021).

[13] Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M. & Suppes, P. A representational similarity analysis of the dynamics of object processing using single-trial eeg classification. *PLOS ONE* **10**, 1–27 (2015).

[14] Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

[15] Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology* **10**, 1–29 (2014).

[16] Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences* **116**, 21854–21863 (2019).

[17] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports* **6**, 27755 (2016).

[18] Kiat, J. E. *et al.* Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Developmental Science* **25**, e13155 (2022).

[19] Kriegeskorte, N. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science* **1**, 417–446 (2015). URL https://doi.org/10.1146/annurev-vision-082114-035447. PMID: 28532370, https://doi.org/10.1146/annurev-vision-082114-035447.

[20] Xu, Y. & Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications* **12**, 2065 (2021).

[21] Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A. & Love, B. C. Measures of neural similarity. *Computational Brain & Behavior* **3**, 369–383 (2020). URL https://doi.org/10.1007/s42113-019-00068-5.

[22] Ramírez, F. M. Orientation encoding and viewpoint invariance in face recognition: Inferring neural properties from large-scale signals. *The Neuroscientist* **24**, 582–608 (2018). URL https://doi.org/10.1177/1073858418769554. PMID: 29855217, https://doi.org/10.1177/1073858418769554.

[23] Schölkopf, B. & Smola, F., A. J.and Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT Press, 2002).

[24] Sahami, M. & Heilman, T. D. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, 377–386 (Association for Computing Machinery, New York, NY, USA, 2006).

[25] Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528 (2011).

[26] Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint: 407007* (2018).

[27] Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience* **35**, 13402–13418 (2015).

[28] Kriegeskorte, N. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience* **3**, 363–373 (2009).

[29] Bowers, J. S. *et al.* Deep problems with neural network models of human vision (2022). URL psyarxiv.com/5zf4s.

[30] Serre, T. Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science* **5**, 399–426 (2019). URL https://doi.org/10.1146/annurev-vision-091718-014951. PMID: 31394043, https://doi.org/10.1146/annurev-vision-091718-014951.

[31] Dujmović, M., Malhotra, G. & Bowers, J. S. What do adversarial images tell us about human vision? *eLife* **9**, e55978 (2020). URL https://doi.org/10.7554/eLife.55978.

[32] Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annual Review of Neuroscience* **42**, 407–432 (2019).

[33] Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nature Reviews Neuroscience* **22**, 703–718 (2021).

[34] Ramírez, F. M. Representational confusion: the plausible consequence of demeaning your data. *bioRxiv* (2017). URL https://www.biorxiv.org/content/early/2017/09/28/195271. https://www.biorxiv.org/content/early/2017/09/28/195271.full.pdf.

[35] Henriksson, L., Khaligh-Razavi, S.-M., Kay, K. & Kriegeskorte, N. Visual representations are dominated by intrinsic fluctuations correlated between areas. *NeuroImage* **114**, 275–286 (2015). URL https://www.sciencedirect.com/science/article/pii/S105381191500316X.

[36] Cai, M. B., Schuck, N. W., Pillow, J. W. & Niv, Y. Representational structure or task structure? bias in neural representational similarity analysis and a bayesian method for reducing bias. *PLOS Computational Biology* **15**, 1–30 (2019). URL https://doi.org/10.1371/journal.pcbi.1006299.

[37] Mitchell, M. & Krakauer, D. C. The Debate Over Understanding in AI's Large Language Models (2022). 2210.13966.

[38] Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020).

[39] Edelman, S. Representation is representation of similarities. *Behavioral and Brain Sciences* **21**, 449–467 (1998).

[40] Block, N. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* **10**, 615–678 (1986).

[41] Fodor, J. & Lepore, E. *Holism: A Shoppers Guide* (Blackwell, Cambridge, 1992).

[42] Hempel, C. G. Problems and changes in the empiricist criterion of meaning. *Revue Internationale de Philosophie* **4**, 41–63 (1950).

[43] Quine, W. V. Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review* **60**, 20–43 (1951).

[44] Firth, J. R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).

[45] Griffiths, T. L. & Steyvers, M. A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (Erlbaum, Hillsdale, NJ, 2002).

[46] Griffiths, T. L., Steyvers, M. & Tenenbaum, J. A probabilistic approach to semantic representation. *Psychological Review* **114**, 211–244 (2007).

[47] Piantadosi, S. & Hill, F. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)* (2022). URL https://openreview.net/forum?id=nRkJEwmZnM.

[48] Fodor, J. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press, Cambridge, 1987).

[49] Haxby, J. V. Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* **62**, 852–855 (2012).

[50] Morcos, A., Raghu, M. & Bengio, S. Insights on representational similarity in neural networks with canonical correlation. In Bengio, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc., 2018). URL https://proceedings.neurips.cc/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

[51] Raghu, M., Gilmer, J., Yosinski, J. & Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

[52] Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 3519–3529 (PMLR, 2019).

[53] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

[54] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[55] Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. Rep. (2009).

[56] Muttenthaler, L. & Hebart, M. N. Thingsvision: A python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in Neuroinformatics* **15**, 679838 (2021).

[57] Nili, H. *et al.* A toolbox for representational similarity analysis. *PLOS Computational Biology* **10**, 1–11 (2014).