Synthese (2022) 200:196
https://doi.org/10.1007/s11229-022-03694-y

ORIGINAL RESEARCH

Mapping representational mechanisms with deep neural networks

Phillip Hintikka Kieval[1]

Received: 19 August 2021 / Accepted: 9 April 2022 / Published online: 5 May 2022

Abstract
The predominance of machine learning based techniques in cognitive neuroscience raises a host of philosophical and methodological concerns. Given the messiness of neural activity, modellers must make choices about how to structure their raw data to make inferences about encoded representations. This leads to a set of standard methodological assumptions about when abstraction is appropriate in neuroscientific practice. Yet, when made uncritically these choices threaten to bias conclusions about phenomena drawn from data. Contact between the practices of multivariate pattern analysis (MVPA) and philosophy of science can help to illuminate the conditions under which we can use artificial neural networks to better understand neural mechanisms. This paper considers a specific technique for MVPA called representational similarity analysis (RSA). I develop a theoretically-informed account of RSA that draws on early connectionist research and work on idealization in the philosophy of science. By bringing a philosophical account of cognitive modelling in conversation with RSA, this paper clarifies the practices of neuroscientists and provides a generalizable framework for using artificial neural networks to study neural mechanisms in the brain.

Keywords Machine learning · Cognitive neuroscience · Connectionism · RSA · Mechanistic explanation · Real patterns · Idealization

1 Introduction

We are experiencing an unprecedented era of explosive progress in Artificial Intelligence (AI) research. This success comes on the back of machine learning systems based on deep neural networks. Over the course of the last decade, deep learning has become by far the most successful approach to AI (Lecun et al., 2015). These networks achieve human-level performance at natural image classification, defeat human

✉ Phillip Hintikka Kieval
pzhk2@cam.ac.uk

[1] Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK

masters of complex strategy games (Silver et al., 2016), and accurately predict the three-dimensional structure of proteins from their amino acid chains (Jumper et al., 2020). Understanding how information is encoded in abstract representations in these systems is crucial for understanding their intelligent capacities and their utility as scientific instruments. This imperative suggests revisiting connectionism, a view that explains intelligence in terms of domain-general capacities to learn abstract representations from low-level perceptual inputs (Hassabis et al., 2017; Buckner et al., 2018).

Connectionists hope to explain cognition using artificial neural networks (hereafter, ANNs) as models (Rumelhart et al., 1986). ANNs consist in large, parallel collections of artificial neurons called units, or nodes, together with weights that measure the strength of connections between nodes. We can view these networks as models of the neurons and synaptic links in the brain at some level of abstraction. Indeed, renewed interest in connectionism within the philosophy of science points out that ANNs can be understood as idealized, multilevel models capable of generating explanatory insight into the mechanisms underlying cognition (Bechtel et al., 1998; Stinson, 2018, 2020). This aligns with the prominent philosophical view that one of the primary explanatory goals of neuroscience is to illuminate the parts and functions of structured mechanisms that give rise to our diverse cognitive capacities (Bechtel, 1998; Cummins, 2000; Machamer et al., 2000; Bechtel & Abrahamsen, 2005; Kaplan & Craver, 2011).

Consonant with these developments in the philosophy of cognitive modelling, mounting empirical evidence from computational cognitive neuroscience suggests that deep neural networks (henceforth, DNNs) are useful tools for learning about regions of interest in the brain (Cichy et al., 2016; Yamins et al., 2014; Yamins & DiCarlo, 2016; Diedrichsen & Kriegeskorte, 2017; Kriegeskorte & Douglas, 2019). Deep learning classifiers now play a pivotal role in multivariate pattern analysis (MVPA) (Haxby, 2012; Kriegeskorte & Diedrichsen, 2019), a set of methods that involve decoding and analyzing information from collected patterns of neural activity (Haxby, 2012). These methods are depicted as probing the abstract representational geometry of the brain (Haxby et al., 2014; Kriegeskorte & Kievit, 2013; Kriegeskorte & Diedrichsen, 2019). Yet, there are complications issuing from the use of DNNs as instruments of measurement in neuroscience. DNNs are ruthless correlation extractors unconstrained by the information that is actually consumed by downstream processes (Carlson et al., 2018; Ritchie et al., 2019). Attempts to justify these measurement methods often reveal a circular reliance on the very same kind of machine learning classifiers. Consider, for example, the observation that the vast majority of variation in complex neural activity can be reduced to a comparably tiny number of explanatory variables. Many take this observation to support the hypotheses that neural representations are sparsely coded. This is taken in turn to support the reliability of data-driven techniques for identifying that neural code. However, the evidence for the former observation primarily derives from the application of such data-driven techniques themselves.

A closely related issue stems from the fact that neuroscientists work with messy data. Neural activity has many degrees of freedom, resulting in very high-dimensional data. With so many degrees of freedom it becomes increasingly difficult to determine which dimensions of variation are functionally relevant for controlling behavior. In other words, how can we distinguish signal from noise? In practice, modelers must

make choices about how to structure their raw data to make inferences about encoded representations. This leads to a set of standard methodological assumptions about when abstraction is appropriate in neuroscientific practice. But when made uncritically these choices threaten to bias conclusions about phenomena drawn from data (Carlson et al., 2018). For example, investigators must select a range of voxels to be analyzed from a given fMRI study to make analysis tractable. But an unprincipled selection process runs the risk of generating spurious correlations in the data. If DNNs are going to do useful work for neuroscience and cognitive psychology, we must confront these methodological barriers head on.

The practices of MVPA will benefit from greater contact with the philosophy of science. In what follows, I consider in detail a specific type of MVPA called representational similarity analysis (RSA) (Kriegeskorte et al., 2008a; Roskies, 2021). RSA is noteworthy for its shared conceptual DNA with earlier connectionist approaches to cognition that use similarity in a semantic space as a measure of representational content (Churchland, 1998; Horgan & Tienson, 1996; Rumelhart et al., 1986). Our recent connectionist ancestors can arm us with a more philosophically informed picture of this methodology. This, alongside work on idealized models in the philosophy science, can illuminate how RSA with goal-driven DNNs can provide legitimate insight into the representational mechanisms in the brain. Ultimately, what I aim to provide is a descriptive account of RSA that shows how and why it sometimes embodies a useful pattern of inference despite its perspicuous limitations.

I begin by introducing the notion of a representational space that forms the basis of RSA. With that on board, I describe the practice of RSA before discussing some of its limitations. I go on to show how the formal machinery of RSA can be traced to earlier approaches to cluster analysis in connectionist modeling. With this analogy in mind, I consider the relationship between RSA and mechanistic explanation. Neuroscientists are concerned to use RSA to study functional mechanisms in the brain, but the exact nature of these inferences are not always explicit. I proceed to argue that RSA can contribute to connectionist-style mechanistic explanations by indirectly linking goal-driven ANNs to target mechanisms in the brain via idealized causal patterns. This indirect route to mechanism through shared causal patterns helps us make sense of the interest-relative idealizations and abstractions present in RSA.

2 What is a representational space?

Representations have long preoccupied philosophers and neuroscientists alike. The most influential philosophical accounts of representational content—especially Fodor (1990), Dretske (1988), and Millikan (1984)—focus primarily on the intentionality of perceptual states. Their hope was to incorporate intelligent behavior into the scientific image of the world in a way that preserved the behavior-guiding role played by the contents of internal states. In seeing that my dog Frasier is now sitting beside me, there is something that my accompanying perceptual state is directed at—my furry, tail-wagging companion. The idea goes that there is some internal state that represents Frasier to the mechanism in my brain responsible for processing sensory inputs and producing the right behavior, namely petting him. First, this involves identifying

physical constituents of a mechanism that we can group into types. It also requires a principled criterion for fixing the meaning of these groups according to the information which they purport to encode. To do this we need to say how a system could manage to carve out features of its environment that are alike in kind and represent them for processing in a physical mechanism. Historically, this has meant showing that representations reliably instantiate the right kind of causal relation with what they represent such that they function to track the relevant kind of phenomena (Dretske, 1988). However, the sense of representation used by neuroscientists tends to be fairly thin and pragmatically oriented (Cao, 2020; Egan, 2020). Representations are typically cashed out in terms of encoded information that can be read out by downstream processes to produce behavior (Kriegeskorte & Diedrichsen, 2019).

So, computational explanations involve identifying representational vehicles and a criterion for fixing their representational content. The vehicles of content are physical particulars over which the network performs computations and which carry information that can be read out by various downstream processes. While individual neurons transform input signals and pass output signals to their neighbors, it would be a mistake to identify them as the vehicles of content. Whether biological or artificial, no two neural networks are the same. Their high degree of variance means a strong neuronal doctrine, which takes individual neurons to be the vehicles of content, rules out the possibility of two networks sharing the same representational content. Rather than identifying individual units as representational vehicles, we typically assume that mental representations are *distributed* across many neurons. Tokened representations are realized by patterns of activity throughout an ensemble of interconnected neurons. Each of these patterns corresponds to a point in an abstract representational space. The semantic content of these representations is then typically analyzed in terms of proximal clusters of activity in an abstract representational space.

Identifying representations from a collection of neural data thus involves a degree of mathematical abstraction right off the bat. Modelers begin by measuring the distributed patterns of activity associated with different experimental conditions. Each token representation will be one such pattern of neural activity distributed across the population. Such a pattern, or activation vector, can be characterized as an n-tuple, where n is the number of neurons in a population (or hidden layer nodes in the case of an ANN). We can conceptualize these activations as points in a geometric space for the purpose of comparing informational content. *State space* refers to an abstract, multidimensional space whose axes are constituted by the possible activation of neurons in a population—or units in the hidden layers of an ANN—such that any pattern of simultaneous activity corresponds to a point in that space. A point in state space corresponds with a token representation. Thus, it will be useful to refer to the state space of a neural network thus characterized as a *representational space.* Many of the practices of cognitive neuroscience involve mapping and analyzing the representational space of a given region of interest in the brain. The various techniques used to carry out this inquiry very often involve the assistance of machine learning classifiers.

3 Representational similarity analysis (RSA)

Over the course of the last twenty years, MVPA has been championed as uncovering the representational structure of the brain. Non-invasive brain imaging techniques like functional magnetic resonance imaging (fMRI) enable scientists to collect population-wide neural recordings from many human subjects. A caveat to this is that fMRI provides only an indirect indicator of neural activity. fMRI measures blood-oxygenation-level-dependent (BOLD) signal, which relates to the ratio of oxygenated to de-oxygenated blood (Passingham & Rowe, 2016). This signal is represented in 3 cubic millimeter regions of tissue called voxels, or volumetric pixels. This means that MVPA depends on measurement channels that are more coarse-grained than individual neurons.

Representational similarity analysis (RSA) is an increasingly prominent encoding model approach to MVPA. Encoding models aim to predict the response patterns of neural activity from descriptions of the experimental conditions (Naselaris et al., 2011; Naselaris & Kay, 2015; Kriegeskorte & Douglas, 2019).[1] RSA provides a framework for comparing the abstract geometries of different representational spaces (Kriegeskorte et al., 2008a; Kriegeskorte & Kievit, 2013). This is achieved by calculating pairwise similarity measures between different patterns of activity and assembling these measures into a two-dimensional matrix called a representational dissimilarity matrix (RDM). An RDM thus summarizes the similarity relations between patterns of activity in response to changes in experimental conditions. Neuroscientists can then use statistical methods to test the similarity of different RDMs generated from distinct representational spaces. This gives a quantitative sense of to what degree two different representational spaces are alike (Kriegeskorte & Kievit, 2013; Roskies, 2021).

Like all methods of MVPA, RSA begins by selecting an array of voxels from whole-brain recordings that correspond to a functional region of interest in the brain. This selection can be seen as a kind of decomposition, whereby the operations that contribute to the overall functioning of a mechanism are associated with various working parts of the whole (Bechtel & Abrahamsen, 2005). RSA proceeds by measuring the activity of these voxels in response to changes in experimental conditions. For example, we might expose subjects to a stimulus set consisting of a series of natural images to study early visual perception. In this case, we will measure the responsiveness of selected voxels to each novel image. This activity across all selected voxels can then be coded as a vector. Such an activation vector captures the distributed activity within the region of interest in response to each novel stimulus. An RDM contains a cell for each unique pair of experimental conditions. Each cell contains a value measuring the similarity between the activation vectors associated with two stimuli according

[1] The encoding methods which are the focus of this paper stand in contrast to decoding methods. Decoding models aim to detect the presence of specific information in the brain by treating a linear classifier as a proxy for the downstream processes that read out information encoded by neural representations from voxelwise data. Encoding models instead work in the same direction as the flow of information in the brain to make comprehensive predictions about the representational space (Naselaris & Kay, 2015; Diedrichsen & Kriegeskorte, 2017). The model itself can be generated by statistical descriptions of participant behaviors, neural data obtained directly from proxy organisms like chimpanzees, or patterns of activity from ANN models trained on some sensory-perceptual task (Kriegeskorte & Douglas, 2019; Martin et al., 2018; Roskies, 2021).

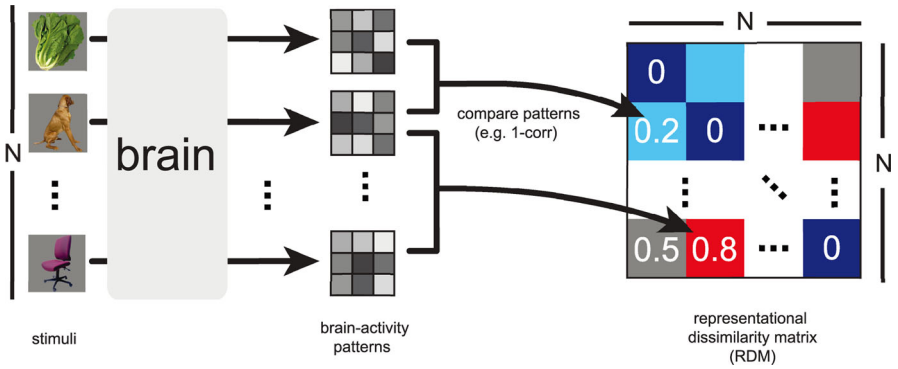


Fig. 1 Computing representational dissimilarity matrices (RDMs) [reproduced from Nili et al. (2014)]

to some chosen metric (e.g Euclidean, correlation, etc.). Entries along the diagonal represent comparisons between identical stimuli and take a value of 0. The value of each off-diagonal entry represents the dissimilarity between patterns in response to two different stimuli. Lower value entries indicate that a pair of stimuli produce more similar responses, while a value of 1 indicates no correlation whatsoever. We then construct the matrix by arranging each stimulus into an order, usually according to an observers' intuitive similarity judgments, and assigning the computed (dis)similarity value to its corresponding cell (Kriegeskorte et al., 2008a). In an experiment with N experimental conditions, this yields an N × N RDM (Fig. 1). Such an RDM provides a two-dimensional map of the similarity relations between a set of activation vectors.

For example, RDMs measuring inferior temporal cortex (IT) neural population responses exhibit a clear block-diagonal structure characteristic of the IT's high performance at object categorization. Predictively adequate neural network models will exhibit a similar block-diagonal structure in their own RDMs. When cells are arranged according to observer similarity judgments, strong structural correlation provides evidence that the activation space implements a representational space (Kriegeskorte et al., 2008a; Roskies, 2021). RDMs can thus be interpreted as a simplified description of the geometry of a given representational space.

Despite the thus far rosy picture of RSA, there is a swath of methodological and philosophical barriers that threaten the prospects of generating genuine explanations with the framework. For instance, the modelers' choice of similarity measure is a decision that has significant implications for how the representational space is reconstructed from the underlying data. It is commonplace to use correlation distance because, unlike standard Euclidean distance, normalization makes it scale invariant.[2] Scale invariant measures make it easier to compare relative distances within different representational spaces because they abstract away from the absolute magnitude of

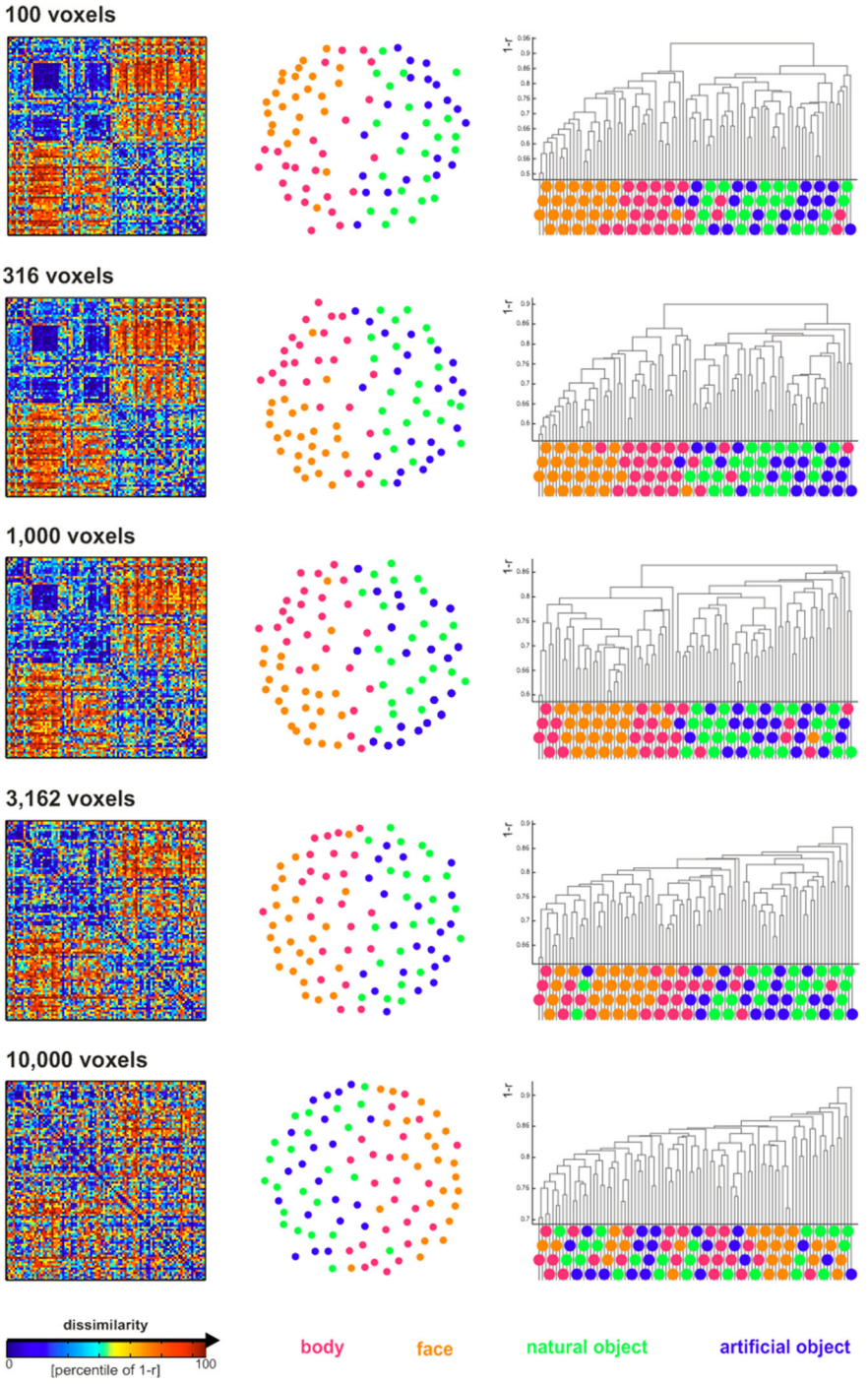
[2] Correlation distance equals 1−Pearson's ρ. Pearson's correlation is a measure of normalized covariance between two variables. It can be calculated by dividing the covariance of two variables by the product of their standard deviations. Correlation can be given a geometric interpretation due to its relation to the angular metric, cosine distance. The cosine of the angle between two vectors can be obtained by normalizing them by their Euclidean distance and then calculating their inner product. The correlation distance between two vectors is equivalent to their cosine distance after subtracting the mean value from each activity pattern.

those distances. However, it remains an open question as to whether this is a principled reason for choosing correlation distance as a similarity measure (Walther et al., 2016; Bobadilla-Suarez et al., 2020). Results from Walther et al. (2016) and Bobadilla-Suarez et al. (2020) both seem to suggest that other measures of similarity better capture the decoding processes used by the brain for certain kinds of stimuli.[3] Moreover, modelers typically opt for some form of noise normalization to better capture the dimensions of variation in the representational space. This typically precedes some kind of data-driven approach to feature selection and subsequent dimensionality reduction. Dimensionality reduction can prove helpful for navigating intractably complex activation spaces and representing only the dimensions that are functionally relevant to a representational space with lower intrinsic dimensionality. However, these data-driven methods are not guaranteed to generate a hypothesis-neutral transition from data to phenomena (Goddard et al., 2018; Carlson et al., 2018). Data-driven dimensionality reduction techniques can reduce the variation found in very high-dimensional recordings to only a few explanatory dimensions, but there is no guarantee that this low-dimensional representation corresponds to the information that plays a functional role in downstream processes. And there are still further idealizing assumptions embedded in these practices. For instance, modelers typically assume that the distribution of noise in voxelwise data is Gaussian. How the modeler chooses to carve up the data affects the structure of the representational space.

Figure 2 illustrates this effect. Kriegeskorte et al. (2008b) used data-driven methods to identify the responsiveness of voxels in the IT cortex to visual stimuli. Sets of voxels were then selected for inclusion in the similarity analysis according to their responsiveness. This marks an implicit decision to treat voxels that are highly responsive to stimuli as functionally relevant to the representational space and to treat less responsive voxels as noise. The number of voxels selected for inclusion impacts the structure and discernibility of patterns in the representational space. As the number of selected voxels increases the categorical structure found in the RDM becomes less distinct. At 10,000 voxels the organization between faces and bodies appears almost completely obliterated. Of course, we have good, theory-driven reasons to suspect that structure is there. Stripping away irrelevant features reveals that structure. However, selecting for certain sets of voxels over others, even when using data-driven methods, runs the risk of biasing results towards a favored hypothesis.

Scientists using RSA thus face various methodological decisions concerning how and when to idealize. Getting on with normal scientific practice requires scientists to make a range of assumptions about when idealizations and abstractions are appropriate. Without careful consideration such assumptions threaten to undermine explanatory findings issuing from RSA.

[3] Admittedly, Bobadilla-Suarez et al. (2020) rely on a familiar but problematic assumption that decodability can stand in as a proxy for information transfer in the brain and that high confidence by a classifier can reliably indicate information gain (2020, pp. 372–373).


100 voxels
316 voxels
1,000 voxels
3,162 voxels
10,000 voxels
1-r
dissimilarity
0
[percentile of 1-r]
100
body
face
natural object
artificial object

◀ **Fig. 2** RDMs (left), MDS arrangements (middle), and dendrogram trees (right) were computed on the same region while varying the number of voxels selected for inclusion. The similarity structure and categorical clustering of human IT varies based on the number of voxels [(reproduced from Kriegeskorte et al. (2008b) supplemental data]

4 The connectionist roots of representational geometry

The combination of quantitative and qualitative analysis of similarity structures within a representational space has many historical precedents in philosophy and cognitive psychology. In fact, all of the tools needed for RSA are already contained within earlier work on connectionist models (see especially Laakso & Cottrell, 2000). Re-examining these connectionist roots will help to clarify the kinds of inferences we can draw from RSA and when those inferences are sound.

Connectionism broadly-construed takes ANNs as idealized models of cognition. An ANN consists in multiple, interconnected layers of units joined together by a pattern of weights which determine the strength of activity passed from one unit to the next. The layers of a network are divided into three classes: an input layer that receives encoded information for processing, an output layer that produces the result of processing, and a hidden layer (or layers) that lies in between. The process of training a simple, supervised classifier network involves introducing a large number of antecedently labeled example inputs and using a supervised algorithm called error backpropagation to fine-tune the associative links between units such that the network learns to successfully generalize to novel inputs.

The de facto solution to the problem of identifying representational content in connectionist models has been to take clusters of distributed activity as the the true vehicles of content (Tiffany, 1999; Rupert, 2001; Gardenfors, 2000; Shea, 2007). Call this proposal the *cluster approach* to content. This approach dates as least as far back as Hopfield (1982) and was brought to prominence in the philosophy of cognitive science by connectionists such as Horgan and Tienson (1996) and Churchland (1998).[4] According to the cluster approach, we can identify clusters of points, or regions in representational space, as corresponding to different representational types (Horgan & Tienson, 1996; Gardenfors, 2000; Shea, 2007). Rather than identifying individual hidden units or the relational structure between individual points as the vehicles of content, the cluster approach identifies clusters of activity in state space as the correct vehicles of content. On this view, representations can be grouped into identical

[4] Churchland (1998) defends a somewhat different version of state space semantics based on relative similarity. This similarity approach takes the overall pattern of simultaneous activation levels across the nodes of a model's hidden layers as the vehicles of content, but holds that a point in a model's state space acquires a specific semantic content as a function of its position relative to all of the other contentful points within that space. So, Churchland replaces the notion of semantic identity with one of relative similarity. Fodor and Lepore (1999) observe, however, that the similarity approach amounts to an unacceptable form of content holism. The content of a particular activation vector depends on its relation to all other contentful points in the space. But the identity of those contentful points each depends on their relations to every other point, and regress threatens to collapse the distinction between tokens and types. It follows that individual representations cannot to be compared between two different models, since comparisons are only possible between entire state spaces. Put another way, if representational content is adjudicated according to similarity in semantic space, then the type-token distinction seemingly breaks down, and there is no principled way of characterizing a model that separates points in state space into types (Fodor, 2000, p. 50).

semantic types in virtue of their falling within separable regions in state space where clusters of neural activity emerge through training. The idea is just that a set of training samples that produces a cluster of points in state space A might likewise produce a cluster of points in state space B. In such a case, two different state space clusters in two different neural networks can represent the very same property to their respective networks.

This approach builds on modeling work by Laakso and Cottrell (2000). This work helped to establish cluster analysis as a procedure for comparing representations between networks with individual differences in weight matrices and architecture. Cluster analysis refers to a family of techniques for measuring the distribution of activation points in state space. Laakso and Cottrell (2000) demonstrated this by comparing clustering patterns in pairs of networks. To do this, they first trained a series of simple ANNs with a variety of different architectures to perform color classification. Each of these were shallow, three-layer networks, but varied in their number of input nodes and number of nodes in hidden layer. Changes at the input layer alter how inputs are encoded by the network, whereas variation in the number of hidden units affects the dimensionality of the the network's state space. Laakso and Cottrell's experiment was thus designed to compare the distribution of activity across networks with different input encoding schemes and dimensionality when sorting through identical data. They ultimately found that their different networks acquired highly correlated arrangements of activity in state space despite the variation in architecture.

To get these results Laakso and Cottrell devised a method to compare distances in state space that would be insensitive to dimensionality. They achieved this by first computing the distances between every unique pair of activation points in each network's respective state space. They then constructed a pairwise matrix of said distance measures for each network. Since each network was trained on the same number of input samples, each matrix would have to be the same size. So, reproducing this procedure for every input/output pair for two different networks results in a pair of matrices with $n(n-1)/2$ unique elements, where n is the number of input samples (2000, p. 57).

Each unique element in a given matrix can be laid out as a vector. Call this a network's *vector coding*. Recall that each of these vectors will have the same number of elements, since each network was trained on the same data. So, to determine the overall similarity of the state space of network A to that of network B, Laakso and Cottrell computed the statistical correlation between their respective vector codings (2000, p. 57). Correlation (using Pearson's ρ) measures the extent to which the values in one data set can predict the values from another. The degree of statistical correlation between the vector codings of two networks gives us sense of whether pairs of input samples which produced proximal activations in the state space of one network produced similarly nearby activations in the state space of another network, and, likewise, whether samples which produced distant pairs in the state space of one network also produced distant pairs in the state space of the other. In short, Laakso and Cottrell's strategy was to construct a distance matrix for each network that contained the same number of elements irrespective of dimensionality, hence enabling the straightforward calculation of statistical correlation. This methodology mirrors that of RSA nearly exactly and laid the groundwork for further refinement by Kriegeskorte et al.

(2008a). In fact, the original paper proposing RSA as a neuroscientific framework, Kriegeskorte et al. (2008a), explicitly cites Laakso and Cottrell as inspiration.

Laakso and Cottrell (2000) argue that their method has a number of advantages. Using statistical correlations in this way purports to achieve scale invariance, since it is insensitive to the magnitude of the distances being compared. We hold true that individuals can share the same beliefs despite presumably different neural architectures. As such, providing a criterion for semantic similarity that eschews the need for an absolute match in the dimensionality of representational space is a crucial desideratum for an adequate state space semantics. If computing the the similarity of distances between points in two spaces is scale invariant, then the cluster approach will satisfy this desideratum.

The direct lineage from connectionist work on cluster analysis can inform our theoretical understanding of RSA. By reframing RSA against the backdrop of connectionist modeling, I argue that we can better understand how RSA can explain neural mechanisms. Traditional connectionist models are taken to comport to the mechanistic account of explanation (Bechtel, 1998; Bechtel & Abrahamsen, 2005; Cummins, 2000; Machamer et al., 2000; Zednik, 2011; Stinson, 2018, 2020). In very broad strokes, mechanistic explanation aims to understand cognitive capacities by isolating the concrete, working parts that generate them. However, the rampant abstraction and idealization present in RSA seems at odds with mechanistic explanation. RSA considers only internal relationships by abstracting away from implementation details and the absolute magnitude of activity measures. In virtue of this, RSA can compare structure in such a way that is invariant across different modalities (Roskies, 2021). Yet, the main thrust of mechanistic explanation depends on relating the behavior of a system to operations performed by the concrete parts of a structured mechanism. I aim to bridge this gap by bringing renewed discussions of connectionist modeling into conversation with RSA. In what follows, I discuss philosophical work on idealization and argue that, much like its connectionist forbears, RSA can sometimes contribute to mechanistic explanations despite such idealizations.

RSA itself is a method for generating idealized representations of a system's representational geometry. This idealization appears to occlude a simple mapping from model to mechanism. But, there is another way of understanding the route to mechanisms that draws on recent work on connectionist modeling (Stinson, 2020). The idealization present in RSA functions to emphasize causal patterns embodied by a mechanism. These causal patterns mediate inferences from neural network models to neural mechanisms. RSA achieves this by connecting neural network models and their targets to shared causal patterns.

5 Models and their targets instantiate causal patterns

This section aims to make explicit the connection between the representational explanations found in RSA and mechanistic explanation. Philosophical work on cognitive modeling emphasizes how models contribute to mechanistic explanations (Kaplan & Craver, 2011; Stinson, 2018, 2020). Neuroscientists themselves gesture towards the goal of understanding functional neural mechanisms, but rarely spell out how RSA

contributes to this kind of explanation (Kriegeskorte & Diedrichsen, 2019). Hence, explicitly specifying how RSA contributes to mechanistic explanations can help to clarify the framework's utility in the face of recent skepticism about representational explanations afforded by computational models in cognitive neuroscience (Carlson et al., 2018; Ritchie et al. 2019; Gessell et al. 2021). I aim to do so by drawing on philosophical work on causal patterns, idealization, and connectionist modeling. Connecting this work to RSA can give us a clearer picture of how it can help explain the functional neural mechanisms that support our cognitive capacities. I argue that idealizations and abstractions present in RSA function to emphasize such causal patterns between mechanisms and their environment. These shared causal patterns mediate inferences from model to mechanism (Stinson, 2020).

This idea of a causal pattern draws upon Andersen's information-theoretic revitalization of Dennettian "real patterns" (see also Dennett, 1991; Potochnik, 2017; Stinson, 2020). Real patterns—as discussed by Andersen (2017) and Potochnik (2017)—refer to *causal* patterns that structure our natural world. For Andersen, causal relationships are informational relationships between patterns instantiated in a rich causal nexus (2017, p. 594). Hence, I will refer to these patterns as causal patterns to avoid confusion. According to Andersen, a causal pattern is one that "can be reliably picked out and tracked through time and which allows one to make predictions that are better than chance" (2017, p. 602). Causal patterns are counterfactually robust: the microphysical state underlying a tokened pattern could have been slightly different, while still tokening the same pattern. The basic idea is that kinds or patterns can be reliably picked out and tracked by information-theoretic means and make useful predictions. This ensures that patterns are not met with jury-rigged kinds. Nevertheless, they remain metaphysically innocuous. Since causal patterns make useful predictions and are stable under counterfactual perturbation, a collection of phenomena that manifest a causal pattern will constitute a robust kind (Stinson, 2020).

To get a handle on what constitutes a causal pattern, consider a digital chess program in which two computational engines are playing each other in a game of chess. The state of this game at any one moment in time can be described as no more than a complex array of pixels, or a bit map. The bit map gives a complete, accurate description of the state of the board at any one instant in the game much like a complete microphysical description of an actual chess board would. In principle, we could compute the entirety of the current board-state using nothing but the bit map. If we know enough details about the algorithms our chess-playing engines implement, then we could use the bit map to accurately predict future board-states. But this would be extremely computationally costly. It is much more efficient to characterize our program at a higher-level of description in terms of chess positions. At this level of description, familiar patterns emerge from the complicated array of flashing pixels. We can identify them as knights, rooks, pawns, and all of the recognizable features that constitute a board-state in a game of chess. Once recognized as a game of chess, enormously more efficient ways of predicting future board-states become available to us with relatively little loss of accuracy (depending on how adept you are at chess). Recognizing these patterns means the difference between computing millions of pixels and merely inferring in your head what is likely to be the best move in an ongoing game of chess. The same idea applies to causal relata. We can make fabulously accurate

predictions by considering the complete microstructure of a physical system, but doing so would be very computationally costly. Instead, we can make predictions about said system on the basis of causal patterns. These predictions may be somewhat less accurate, since information is lost in the move to the level of patterns. But, we are more than compensated for this loss with subsequent improvements in efficiency. In short, we sacrifice a small degree of fidelity for large gains in efficiency. A causal pattern just is any such description that is more efficient than a bit map and facilitates prediction.

As noted above, this notion of causal patterns can help us make sense of the relationship between idealized models and mechanistic explanations. Idealization has enjoyed significant philosophical attention, especially in the context of scientific models (Cartwright, 1994; Morgan & Morrison, 1999; Morgan, 2002, 2003; Weisberg, 2007, 2013; Rohwer & Rice, 2013). Here it is useful to draw on Potochnik's extensive account of the relationship between patterns and idealization in science. She describes how the practice of science by limited humans in a complex world leads to widespread idealization. Potochnik (2017) argues that scientific enterprise can be understood as a search for causal patterns in the face of "causal complexity". Causal complexity refers to the fact that phenomena under investigation are causally influenced by a multitude of factors beyond the variables targeted by the investigation, including by controls implemented to highlight the targeted variables and many other influences (Potochnik, 2017, p. 35). Science aims to represent patterns embodied by causally complex phenomena. This is because patterns play a central role in promoting understanding and manipulability.

Causal complexity also means that phenomena can embody many patterns. The more causal influences there are on a target phenomenon, the more variables there are to produce patterns. Scientists thus have a choice not only about what phenomena should be the targets of study, but also about which patterns should be the focus of their study. Just as competing pragmatic and epistemic considerations inform the design decisions specifying how a single pattern is described, human interests inform which patterns will be useful for making scientific progress. Causal patterns that are of interest to scientists tend to be simple and general enough to facilitate understanding, but fine-grained enough to provide the basis for manipulability.[5]

Specifying patterns that are productive for science motivates idealization. Idealizations are assumptions made without regard for whether they are true and often with full knowledge that they are false (Cartwright, 1994; Potochnik, 2017, p. 42). Physics often assumes frictionless planes even though no such planes exists. But idealizations in science often take subtler forms. A linear regression model draws a curve that purports to represent the relationship between variables and treats divergence from this curve as error or noise. This is done with full knowledge that divergence in the data also represents observed facts. Cartwright (1983) calls this *representation as-if*. Idealizations of this kind interpolate and extrapolate beyond the data to identify the pattern they instantiate. Elgin (2004) suggests that these "felicitous falsehoods" play

[5] The emphasis on manipulability is meant to capture that information-theoretic causal patterns are by and large causal patterns (Woodward, 2003). Causation, however, is a metaphysically fraught topic. Those unfriendly to manipulability accounts of causation should still feel free to regard causal patterns as both cognitively and pragmatically valuable for their role in facilitating understanding and manipulability.

a cognitively valuable role in science. They "impose an order on things, highlight certain aspects of the phenomena, reveal connections, patterns and discrepancies, and make possible insights that we could not otherwise obtain" (Elgin, 2004, p. 127). The positive content of idealizations is to center the relevance of some causal pattern. Potochnik suggests that "idealizations contribute to understanding by representing as-if to the end of depicting a causal pattern, thereby highlighting certain aspects of that phenomenon (to the exclusion of others) and revealing connections with other, possibly disparate phenomena that embody the same pattern or, in some cases, that are exceptions to that pattern" (Potochnik, 2017, p. 97). So idealizations, understood as deliberate false assumptions, can contribute to successful scientific representations by highlighting causal patterns in what they are intended to represent.

This picture also distinguishes idealizations from abstractions. The former intentionally represents a target system as different than it actually is, while the latter merely omits certain details. Abstractions represent by ignoring some features of the target phenomenon that are inconsequential for the representation. This generates a picture of idealizations and abstractions as distinct but compatible practices of scientific investigation. These practices often become intertwined in scientific models.

The above is abundantly clear in neurocognitive modeling. Computational models are built out of transistors, not cells. In comparing a target system, like the visual pathway, to a computational model, scientists apply abstractions that coarse-grain the target system and its computational model. These abstractions relinquish certain details in order to meaningfully compare the abstract dynamics of target system and model. The enduring details are those scientists take to be functionally relevant for providing an adequate explanation of the phenomena of interest. Simply describing the system in terms of a state space constitutes one such kind of abstraction. But with complex biological systems scientists often need to make difficult choices about what to measure and what features of the abstract dynamics are functionally relevant. These choices constitute, either implicitly or explicitly, a commitment to some degree of abstraction as adequate for explaining their target phenomenon. Precisely how much abstraction qualifies as empirically or predictively adequate should be informed by principled theoretical and pragmatic reasons. Similarly, linear transformations allow for a class of systems with the same kinds of underlying components to be considered a single explanatory target, even when the actual target systems differ in some substantial way (Cao & Yamins, unpublished, p. 10). Such transformations are used to address the fact that there is no one-to-one mapping between neurons in different brains. The basic idea is that we can define a transform class that consists of a set of linear maps between populations of neural activity in different individuals. This transform class posits that activity of any neuron in a given region of one individual can be reproduced by a linear combination of neural activity in the corresponding region of another individual.[6]

RSA involves a similar combination of abstraction and idealization. For instance, feature selection idealizes by treating a subset of a neural population as functionally relevant for a given task. Similarly, the invocation of representational geometry marks

[6] Stated more formally, this means that for all input stimuli x, the neuronal responses to x by neuron i in individual T can be written as $n_i^T(x) = \sum_{j=1}^{M} a_{ji} n_j^S(x)$ where a_{ji} are constants that represent the contribution each source neuron makes to replicating the activity of the target neuron (Cao & Yamins, unpublished, p. 7).

a common abstraction between models and their targets. Yet, this picture of idealized cognitive models seems at odds with the goals of mechanistic explanation. On the standard model-mechanism-mapping (3M) account, for abstract dynamical models to have explanatory force in systems and cognitive neuroscience there must be a plausible mapping between elements in a model and the parts of the mechanism they represent (Kaplan & Craver, 2011). Instead, RSA involves mappings between models and their targets that are mediated by a common abstraction to representational space. But, this appears to conflict with the original point of 3M, which is to connect abstract dynamics to particular, concrete mechanisms. How, then, can we expect to make inferences about mechanisms using this idealized framework?

I think this question is best answered by considering the role of causal patterns in mediating inferences to mechanisms. Along these lines, I contend RSA is better positioned as a specific application of Catherine Stinson's account of model inferences in cognitive science (2020, p. 602). Stinson (2020) gives an account of models on which inferences about a target system are drawn from connectionist models indirectly via the kinds of phenomena that both model and target exemplify. Her concept of "kinds" is equally well captured by the notion of causal patterns I am working with here. This indirect route to mechanism helps to make sense of the role of idealization in neuroscientific modelling, since these models aim to capture the characteristic causal patterns associated with a kind of phenomenon. On this view, clustering patterns in an RDM are idealized representations of causal patterns that enable us to efficiently predict the behavior of a system. When it makes statistical comparisons of different representational spaces, RSA functions as a test of whether two different systems (i.e. a model and its target) instantiate the same causal pattern. When a model and target embody the same pattern, we can have confidence that the consequences produced by manipulating the model can be brought to bear on our understanding of the target mechanism.

Tools like RSA can thus provide the confirmatory evidence needed to link a model and target system by highlighting shared causal patterns. The visual pathway and a neural network model both exemplify a particular cognitive capacity, namely early visual perception. The goal-driven nature of visual perception gives rise to certain causal patterns—such as the formation of separable clusters of activity—that our models make salient. We can use these causal patterns to establish an indirect link between neural mechanisms and models designed to solve similar tasks. This is because systems that need to solve similar tasks will tend to instantiate the same kinds of causal patterns. The idea here is simply that, under the right circumstances, function constrains mechanisms. For instance, visual perception places functional and architectural constraints on neural mechanisms. The informational structure of the world is far too complex to represented at the "pixel" level (Dennett, 1991). Such computational costliness begets representational compression as a driving force on neural mechanisms (Gluck & Myers, 2001; Buckner, forthcoming). The high-dimensional encoding scheme of the brain suggests that learning may require domain-general inductive biases that impose structure on the space of possible activity (Poldrack, 2020). Error-correction learning tends to economize representational resources by treating stable clusters of activity patterns that predict the same output as identical. These clustering patterns support generalization to the extent that they conform to the informational structure of the

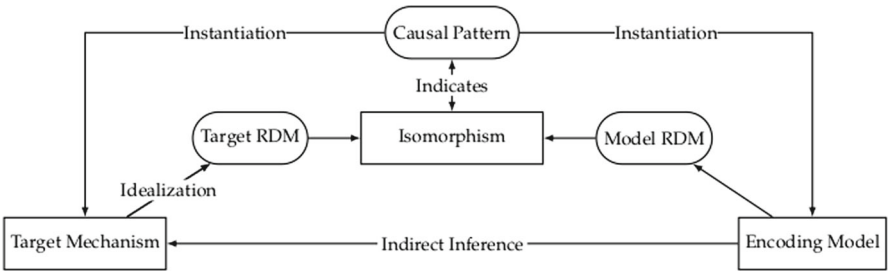


Fig. 3 RSA gives us evidence that an encoding model instantiates the same kind of causal pattern as its target. These shared causal patterns facilitate indirect inferences about the target

environment. RSA can help us to highlight these patterns, however idealized, in the model and compare them to their target system. When we can use RSA to predict the behavior of the target from a model, we get evidence that they instantiate the same causal pattern (Fig. 3).

We can thus use neural networks as minimal models to make inferences about the mechanisms responsible for early vision because both instantiate the relevant causal patterns. RSA provides a method of establishing this link. We thus need abstraction and idealization to make these patterns salient. For a fully-runnable computational model, this merely requires reading off activity of each unit in the model for each stimulus response and plotting a trajectory through representational space with those values. For the target system, we may have to do more work, such as defining a transform class to represent the contributions of multiple source neurons as identical to the contribution of a single model unit. The result of this transformation represents the state space of the target system as if it was isomorphic to the model's state space. Such a tranformation thus constitutes a kind of idealization that operates in tandem with prior abstractions to represent a target causal pattern.

By shifting focus to causal patterns, we can better understand how neuroscientists can make inferences about neural mechanisms via the manipulation of idealized models. I argue that this should be seen as the explanatory goal of RSA. Moreover, what causal pattern occupies the focal point of a particular scientific explanation is, to some degree, sensitive to the interests and goals of scientists. Therefore, it should not come as a surprise that even more data-driven methods of decomposing a high-dimensional state space are not hypothesis-neutral, nor should this be seen as a devastating flaw. Causal patterns that are of fundamental interest to the explanatory goals of scientists already constrain the hypothesis-space. This is legitimate so long as those interests are sufficiently well-motivated by empirical and theoretical considerations. Carlson et al. (2018) make a similar remark that hypothesis-driven approaches that use dimensionality reduction can be defensible, "so long as they are carefully constrained" (Carlson et al., 2018, p. 95). As long as these conditions are satisfied, different ways of decomposing state space in order to make causal patterns salient to scientists are sound methods of testing whether a target phenomenon really does embody the causal pattern of interest.

Moreover, by emphasizing mechanistic rather than representational explanations, we can somewhat sidestep skeptical worries about the latter. Even in cases where it seems like the specific content of representations is radically underdetermined by the available evidence (Gessell et al., 2021), we can still use RSA to get evidence about the relationship between a model and a target system. This is because our goal is to test whether a model and its target both instantiate some idealized causal pattern. When our aim is broadly to understand functional mechanisms, we can do this without attending so closely to problems concerning representational content. Put another way, we can think of RSA as a means of reducing "link uncertainty," which occurs when there is a dearth of evidence supporting the link between a model and its target (Sullivan, 2020). None of this is to deny the fruitfulness of more straightforward representational explanations using RSA. Rather, casting RSA in terms of testing whether different systems embody causal patterns sheds light on how RSA can play a role in mechanistic explanations even when representational content is underdetermined by the data.

6 Inferring neural mechanisms with goal-driven models via causal patterns

ANNs provide us with idealized models of biological neural pathways. These models explain neural mechanisms indirectly. They do so by instantiating a shared causal pattern between sensory mechanism and environment. Modern deep neural networks (DNNs) are optimized to solve the same sensory categorization tasks faced by the brain. This "goal-driven" approach turns the search for neuroscientific explanations into an optimization problem, where the goal is to maximize the accuracy of predictively adequate models (Yamins & DiCarlo, 2016). Recent work using the goal-driven approach has uncovered a surprising fact about deep convolutional neural networks (DCNNs). Models optimized merely to classify images predict spiking responses in the highest level of the ventral stream, the inferior temporal cortex (IT) (Yamins et al., 2014; Cichy et al., 2016). That such task-optimized models manage to predict something about the brain supports the notion that these neural networks and the primate visual pathway embody the same causal patterns.

RSA provides a method for drawing out these indirect inferences about representational mechanisms in the brain using task-optimized DNNs as idealized models. RSA gives us evidence that model and target both instantiate a hypothetical causal pattern. With this link established, neuroscientists can then manipulate these goal-driven models to make inferences about neural mechanisms. A clear proof of concept can be found in Khaligh-Razavi and Kriegeskorte (2014). Khaligh-Razavi and Kriegeskorte (2014) analyzed brain responses in both monkey IT and human IT for a set of color images of objects spanning a range of animate and inanimate categories. They then used RDMs to compare these representations to those generated by 37 different computational models of varying designs. To measure the strength of clustering they created ten category-cluster RDMs as predictors, which they fit to each IT and model RDM (Fig. 5). These grouped the set of experimental conditions according to a number of intuitive categories. The category-clusters represented animate, inanimate, face, human face, non-human face, body, human body, non-human body, natural inanimate, and artifi-

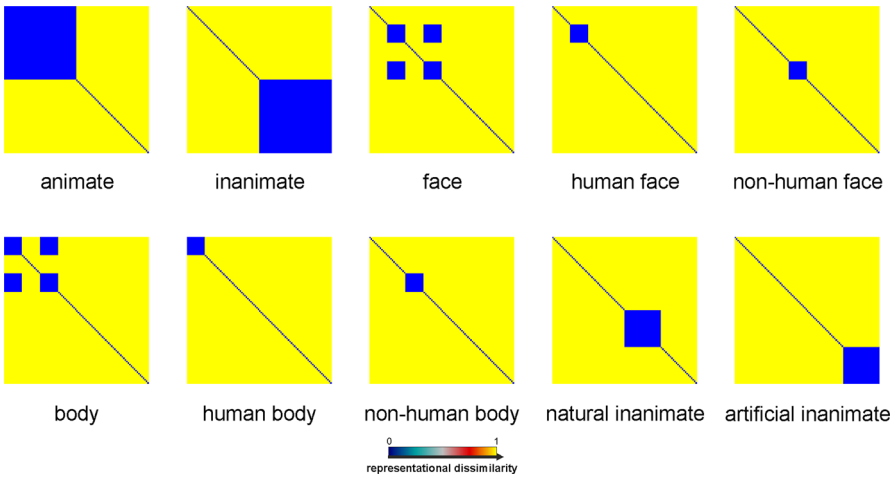


Fig. 4 Khaligh-Razavi and Kriegeskorte (2014) created ten different category-cluster RDMs as predictors of clustering. These prediction RDMs were then fit to each computational model using a linear regression to model the semantic structure of the their representations [(reproduced from Khaligh-Razavi and Kriegeskorte (2014)]

cial inanimate (Khaligh-Razavi & Kriegeskorte, 2014). Though models designed to emulate the structure of the ventral stream (such as HMAX and VisNet) were included among the 37 computational models, they found that these were outperformed in predictive accuracy by a task-optimized DCNN. The representations generated by the supervised DCNN best predicted the the category clustering found in the IT RDMs (Fig. 4).

One way of interpreting this procedure is that the modelers identified informational patterns in their experimental conditions and constructed a model of a conceptual space around these patterns to generate the category-cluster RDMs. By fitting this categorical model to an encoding model, we can identify the causal patterns that arise as plausible candidates for the content of representations. The modelers then used these patterns to predict the representational structure of IT RDMs. What we find is that—much like the primate visual system—higher levels of processing in the DCNN begin to approximate the same patterns. Each subsequent layer represents and processes higher-level properties of the input stimuli with greater tolerance for noise and nuisance variation than the layers preceding it. This procedure clearly concerns the sense of patterns discussed in Sect. 6. However, one might be interested in the second sense of causal patterns. Hence, we might want to abstract away from the specific contents of clusters and instead consider the how particular architectures in the brain structure activity in a way that supports function (Fig. 5).

Comparing brain and model RDMs establishes a kind of second-order isomorphism between the representational geometries produced by both model and brain (Kriegeskorte et al., 2008a; Kriegeskorte & Kievit, 2013; Roskies, 2021). This suggests that both our computational model and its target system instantiate the same causal patterns between a representational mechanism and its environment. This allows these systems to efficiently identify new stimulus conditions and generalize successful per-

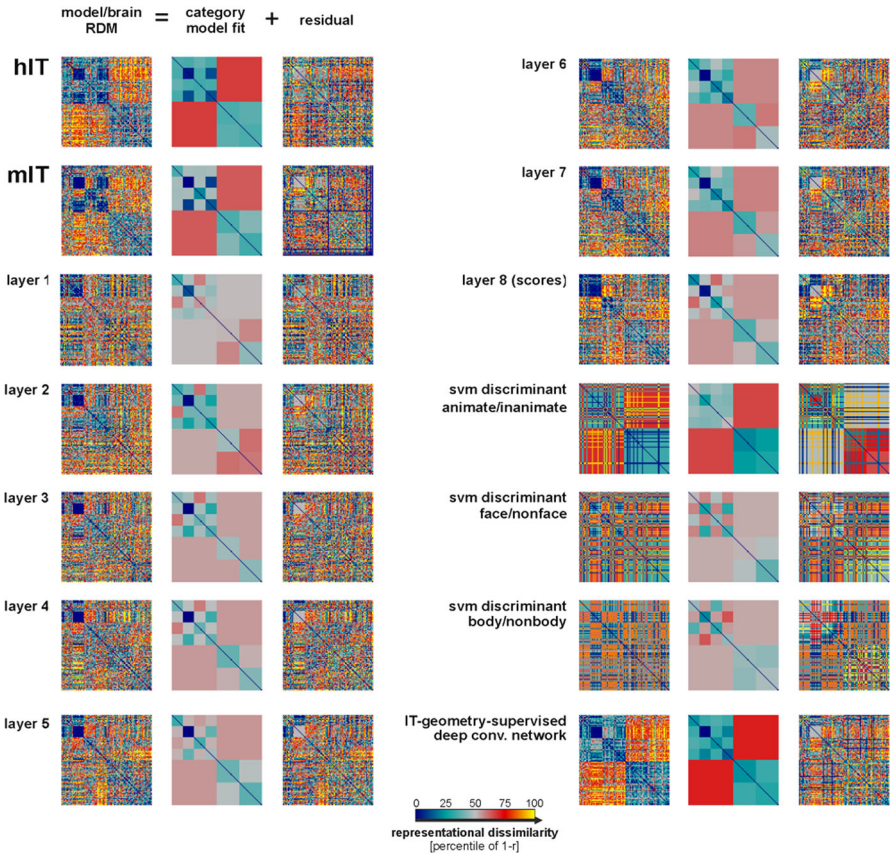


Fig. 5 Category-cluster analysis of a supervised DCNN model reveals that a similar representational structure to human and monkey IT neuronal population responses emerges across the various layers of the model. The final weighted combination of layers (bottom right) shows a similar clustering structure to that of the hIT and mIT [reproduced from Khaligh-Razavi and Kriegeskorte (2014)]

formance to these conditions. If this interpretation is right, then DCNNs turn out to be predictive of neural activity in the primate visual system precisely because they both instantiate the same causal patterns between mechanism and stimulus conditions. Establishing this fact allows the modelers to draw out inferences about the constitutive mechanisms responsible for early vision and the structure of representations acquired by such mechanisms from a fully-runnable DCNN.

More recently, Martin et al. (2018) provide an instructive use of RSA that integrates multiple different regions of interest into a single study. They aimed to understand how the brain integrates clusters of low-level perceptual features with high-level conceptual properties. Many ordinary objects—for instance, hairdryers and guns—tend to possess highly confusable conjunctions of perceptual features, while having radically diverging sets of functional and conceptual properties. Successful behavior requires that we can effortlessly distinguish between these objects. An explanation of object recognition and semantic memory must provide insight as to how the brain integrates these different

kinds of information. Using RSA, Martin et al. (2018) compared multiple behaviorally derived statistical models, neural network models, and fMRI data to identify candidate regions of interests where this conceptual integration likely takes place. Martin et al. asked volunteers to answer questions about different objects while lying inside a neuroimaging machine. These questions concerned both the appearance of objects as well as abstract, conceptual knowledge about them. From participants' responses they constructed behavior-based visual and conceptual RDMs. These models captured visual and conceptual similarity of object concepts separately. They found that these behavior-based RDMs were not significantly correlated, ensuring that perceptual and conceptual semantic dimensions would not be confounded (Martin et al., 2018, pp. 4–5). Additionally, they constructed corresponding brain-based RDMs generated from activity retrieved from voxelwise fMRI data and a conceptual neural network model RDM constructed from activation vectors derived from a fully-runnable neural network model. The brain-based RDMs used multi-voxel activity patterns obtained from multiple regions of interest selected based on empirical evidence linking them to different functional roles in visual and conceptual object recognition. These regions included perirhinal cortex (PRC), the temporal pole, parahippocampal cortex, and lateral occipital cortex. The neural network model RDM was generated from a word2vec-based natural language model (Mikolov et al., 2013), which mapped 3 million words to 300 vectors in a high-dimensional feature space. This word2vec-based model RDM turned out to be significantly correlated with the behavior-based conceptual RDM (Martin et al. 2018, p. 6).

Martin et al. (2018) then conducted second-order RSA to compare visual and conceptual behavior-based RDMs with their respective brain-based counterparts. The results of their study showed several important findings. Their approach suggested that context-responsive visual, conceptual, and integrated visual-conceptual semantic content are represented in distinct similarity codes across several different regions of interest. Most notably, their results support the notion that activity patterns in the PRC represent both visual and conceptual similarity of objects when participants made judgements about either the visual or conceptual features that characterized an object concept. In other words, activity in the PRC captures both the visual similarity of hairdryers and guns and the conceptual similarity of hairdryers and hairbrushes regardless of task context (Martin et al., 2018, pp. 10–11).

The studies above suggest a promising avenue for conceptualizing RSA against the backdrop of philosophically informed modelling principles. What Martin et al. (2018) appear to be after is evidence about the structure of mechanisms responsible for complex cognitive functions. We can thus think of RSA as a hypothesis-driven search for causal patterns with the end goal of uncovering new insights concerning the constitutive mechanisms underpinning our cognitive capacities. The framework I have in mind involves a hypothesis about a target system, a neural network model, and a shared causal pattern. It begins by identifying some cognitive phenomenon and its putative constitutive mechanism. This mechanism—a region of interest in the brain, for example—picks out the target system. Cognitive phenomena of interest are identified by the kinds of robust, generalizable causal patterns that obtain between a target system and its environment. Identified causal patterns help us isolate hypotheses about how the target represents the informational structure of a neural task. Together, these

elements comprise a hypothesis about some aspect of intelligent behavior. Borrowing an example from Martin et al. (2018), modelers might hypothesize that PRC is responsible for the integration of visual and conceptual semantic information. Neural network models can then be selected or designed using a goal-driven approach. These models should be optimized to perform the same kinds of tasks solved by the targets in the brain with goal of instantiating a shared causal pattern with the target. That these models do instantiate the same pattern is established via the statistical tools of RSA. Moreover, this might motivate novel architectures and combinations of multiple networks to tackle increasingly complex, high-level tasks. In the case above, this might involve a model of visual perception like a DCNN combined with a word2vec-based natural language model, and, perhaps, a third network that integrates inputs from both networks to arrive at a unified object concept.

A further advantage arises from the fact that—unlike other kinds of non-human surrogates—neural network models are fully-runnable. This makes manipulating and analyzing their structure more tractable. Insofar as neural network models and their targets instantiate shared causal patterns, we can make predictions and inductive inferences about how a target system approximates the informational structure of its environment. When clusters of activity in a model are predictive of some object category, the presence of that cluster provides evidence that the model has acquired a robust categorical representation. If we can then establish that their model embodies the same causal pattern as the target system, the modelers can then use RSA to collect evidence about the presence of category representations acquired by the target system. Once a model is established via RSA, we can manipulate or perturb the model and observe the downstream effects. This can provide us with indirect insight into how the physical structure of neural mechanisms produce the representational geometries that support robust performance.

Finally, by shifting focus to illuminating mechanisms, we can somewhat sidestep worries about the underdetermination of semantic content by the available evidence (Ritchie et al., 2019; Gessell et al., 2021). Though we are still concerned to understand the content of representations, establishing a high degree of predictive success can be enough to establish that a model instantiates a shared causal pattern. This approach enables computational cognitive neuroscientists to collect confirmatory evidence about the mechanisms hypothesized to be responsible for cognition in virtue of the shared causal patterns instantiated by neural network models. RSA provides a formal toolbox for highlighting salient patterns and establishing task-optimized neural networks as models of cognitive mechanisms. Modelers can then use this framework to draw out inductive inferences about the mechanisms responsible for various cognitive phenomena, generate predictions about target systems from established models, and correct for model errors given what is known about the target system. We can begin modeling more complex, high-level phenomena by combining diverse architectures operating on multiple sensory modalities with the ultimate goal of shedding light on the nature of our intelligent capacities.

7 Conclusion

I have just suggested a framework for understanding the practices of RSA in light of philosophically informed modeling principles. This should help us to make sense of the prevalent idealization that RSA studies rely on. The upshot of this is that it motivates a rigorous, empirical approach to studying localized aspects of cognition with connectionist models. If we want to understand the various mechanisms underlying cognition, we should begin by isolating regions of the brain which might instantiate those mechanisms. We can use RSA as a formal tool to establish similarity between a model and the region of interest in the brain in virtue of some shared causal pattern. This also motivates a more exploratory approach to AI-driven neuroscience. By testing models with varying designs and degrees of neurophysiological inspiration we can begin to uncover the minimal features of the underlying mechanism that are essential for reproducing the cognitive phenomenon of interest. Through such an iterative process we approach a genuinely explanatory mechanistic model of that cognitive phenomenon.

Funding Funding was provided by Gates Cambridge Trust.

References

Andersen, H. K. (2017). Patterns, information, and causation. *The Journal of Philosophy, 114*(11), 592–622.
Bechtel, W. (1998). Representations and cognitive explanations: Assessing the Dynamicist's challenge in cognitive science. *Cognitive Science, 22*(3), 295–318.
Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 421–441.
Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2020). Measures of neural similarity. *Computational Brain & Behavior, 3*(4), 369–383.
Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.
Buckner, C. (forthcoming). *A forward-looking theory of content*. Ergo.
Cao, R. (2020). Computational explanations and neural coding. In *The Routledge Handbook of the Computational Mind*, (pp. 283–296).
Cao, R. & Yamins, D. (unpublished). Making sense of mechanism: How neural network models can explain brain function.
Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage, 180*, 88–100.
Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.
Cartwright, N. (1994). *Nature's capacities and their measurement*. Oxford University Press.
Churchland, P. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *The Journal of Philosophy, 95*(1), 5–32.

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*, 1–13.

Cummins, R. (2000). How does it work? Versus What are the laws? Two conceptions of psychological explanation. In *Explanation and cognition*, (pp. 117–144). The MIT Press.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology, 13*(4), e1005508.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. The MIT Press.

Egan, F. (2020). a deflationary account of mental representation. In J. Smortchkova, K. Dołrega, & T. Schlicht (Eds.), *What are mental representaions?* (pp. 26–53). Oxford University Press.

Elgin, C. Z. (2004). True enough. *Philosophical Issues, 14*(1), 113–131.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT Press.

Fodor, J., & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *The Journal of Philosophy, 96*(8), 381–403.

Fodor, J. A. (1990). *A theory of content and other essays*. MIT Press.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press.

Gessell, B., Geib, B., & De Brigard, F. (2021). Multivariate pattern analysis and the search for neural representations. *Synthese*, (0123456789).

Gluck, M. A. & Myers, C. E. (2001). *Gateway to memory—Introduction to neural network modeling of the hippocampus and learning. Issues in clinical and cognitive neuropsychology*. The MIT Press.

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., & Carlson, T. A. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *NeuroImage, 180*(2017), 41–67.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage, 62*(2), 852–855.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*, 435–456.

Hopfield, J. J. (1982). *Neural networks and physical systems with emergent collective computational abilities (associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)*. Technical report.

Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology*. MIT Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Stanislav, N., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Steinegger, M., Pacholska, M., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2020). High accuracy protein structure prediction using deep learning. In *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*.

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.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11), e1003915.

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

Kriegeskorte, N., & Douglas, P. K. (2019). Interpreting encoding and decoding models encoding and decoding: Concepts with caveats HHS Public Access. *Current Opinion in Neurobiology, 55*, 167–179.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401–412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 1–28.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of Man and Monkey. *Neuron, 60*(6), 1126–1141.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology, 13*(1), 47–76.
Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife, 7*, 1–29.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013—Workshop Track Proceedings*, (pp. 1–12).
Millikan, R. G. (1984). *Language, thought, and other biological categories*. MIT Press.
Morgan, M. S. (2002). *Model experiments and models in experiments*.
Morgan, M. S. (2003). Experiments without material intervention: model experiments, virtual experiments, and virtually experiments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 216–235). University of Pittsburgh.
Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge University Press.
Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences, 19*(10), 551–554.
Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage, 56*(2), 400–410.
Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4), e1003553.
Passingham, R. E., & Rowe, J. B. (2016). *A short guide to brain imaging: The neuroscience of human cognition*. Oxford University Press.
Poldrack, R. A. (2020). The physics of representation. *Synthese*.
Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science, 70*(2), 581–607.
Rohwer, Y., & Rice, C. (2013). Hypothetical pattern idealization and explanatory models. *Philosophy of Science, 80*(3), 334–355.
Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: Proxy vehicles and provisional representations. *Synthese*.
Rumelhart, D. E., McClelland, J. L., & PDP Research Group, C., editors (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: foundations*, (Vol. 1). MIT Press.
Rupert, R. D. (2001). Coining terms in the language of thought: Innateness, emergence, and the lot of Cummins's argument against the causal theory of mental content. *The Journal of Philosophy, 98*(10), 499.
Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind and Language, 22*(3), 246–269.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.
Stinson, C. (2018). Explanation and connectionist models. In Sprevak, M. and Colombo, M., (Eds.), *The Routledge Handbook of the Computational Mind*.
Stinson, C. (2020). From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science, 2019*, 1–38.
Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.
Tiffany, E. (1999). Semantics San Diego style. *The Journal of Philosophy, 96*(8), 416.
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.
Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy, 104*(12), 639–659.
Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

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

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science, 78*(2), 238–263.