Using deep neural networks and similarity metrics to predict and control brain responses

Bojana Grujičić[1,2,3] & Phyllis Illari[3]

[1]Max Planck School of Cognition, Germany
[2]Humboldt-Universität zu Berlin, Berlin School of Mind and Brain, Germany
[3]University College London, Department of Science and Technology Studies, UK

Abstract

In the last ten years there has been an increase in using artificial neural networks to model brain mechanisms, giving rise to a deep learning revolution in neuroscience. This chapter focuses on the ways convolutional deep neural networks (DCNNs) have been used in visual neuroscience. A particular challenge in this developing field is the measurement of similarity between DCNNs and the brain. We survey similarity measures neuroscientists use, and analyse their merit for the goals of causal explanation, prediction and control. In particular, we focus on two recent intervention-based methods of comparing DCNNs and the brain that are based on linear mapping (Bashivan et al., 2019, Sexton and Love, 2022), and analyse whether this is an improvement. While we conclude explanation has not been reached for reasons of underdetermination, progress has been made with regards to prediction and control.

1 Introduction

Measuring things is a much more complicated task than it appears when we use familiar long-embedded methods. Chang (2004) is an excellent lesson in this, as the book examines the long and complex history of temperature, showing how we came to have what we regard as simple and reliable tools to measure temperature, like thermometers. More recent work on measurement can be found in Tal (2020, 2013).

There are now many and varied attempts to causally explain and understand the brain, using many different tools, techniques, and models. In this chapter we will examine the use of Artificial Neural Networks (ANNs) to explain object recognition by modeling the brain mechanism responsible for it, looking at their reliance on various metrics to establish that correspondence, and probing very recent work that attempts to go beyond previous methods. What kind of metrics are needed to establish claims of causal, mechanistic similarity of ANNs and brains in object recognition is the central issue of this chapter.

To relate this specifically to the five scientific problems of causality identified by Illari and Russo (2014), i.e. inference, prediction, explanation, control and reasoning, we will directly address explanation, focusing on the key role of similarity metrics in explanation in this field. However, the complexity of what is addressed ranges more broadly to questions of reasoning that are also involved when the scientists talk about ‘causal understanding’ and also inference, as

some of this work is aimed at identifying causal mechanisms. These wider aims still rely strongly on metrics.

We will first introduce ANNs in section 2, noting some very strong claims concerning how they advance explanation and understanding of the brain in visual neuroscience. In section 3 we dig into how thoroughly dependent that use of ANNs is on claims of similarity between ANNs and brains, relying on similarity metrics, particularly Representational Similarity Analysis (RSA) and Linear Mapping (LM). We discuss why RSA and LM are insufficient metrics to establish similarity of ANN and brain mechanisms. In sections 4 and 5 we examine in detail two recent papers which raise concerns about the limitations of RSA and LM, and attempt to go further by introducing intervention-based comparisons. These are the only two such studies we are aware of, which also go unacknowledged in recent criticisms of RSA and LM-based comparisons of models and the brain by Bowers et al. (2022). While they suggest turning to psychological evidence to arbitrate between models, we explore this other avenue.

Both of the papers we discuss in sections 4 and 5 are based on studies probing a difficult to explore space, albeit in very different ways. In some sense or other, both are looking for possible *parts* in this space: Bashivan et al. (2019), the first study, by making novel images that successfully drive neurons, while the second, Sexton and Love (2022) ‘swap in’ brain data for the relevant layer of the ANN, while retaining the ANN’s functionality. We will analyze their efforts, agreeing that they have indeed gone beyond RSA and LM whose shortcomings have recently become more salient (Bowers et al., 2022, Grujičić, 2023), thus bringing us one step

closer to causal explanation and understanding of object recognition. Intriguingly, Bashivan et al. (2019) seem to begin from parts, while Sexton and Love (2022) begin from operations, making a comparison of the two approaches more illuminating than studying just one.

Ultimately, for reasons of underdetermination stemming from contradictory localisations of operations into parts that we discuss towards the end, we conclude explanation hasn't quite been reached although progress has been made with regards to prediction and control. With regard to the scientific problems of causality above, however, we can see that going beyond prediction is crucial to causal explanation here. Indeed, control of some kind or other is also crucial to the evidence for causal mechanistic explanation. Note, though, that while we talk about interventions, in this space the idea is linked to experimental practices, and not to the more narrowly defined notion used by Woodward (2003).

2 Using ANNs to understand the brain

The history of relating artifacts to human beings to help us understand ourselves is genuinely ancient. Reasoning about analogies between artifacts and the human mind stretch at least as far back as Plato's Chariot in his *Phaedrus*, where he uses the idea of conflict between the black horse and the white horse drawing the chariot to model conflict in the mind, or *psyche*. It should not be surprising that, as soon as computing machinery was in practical use, ideas from it were turned to understanding the human mind and vice versa (Turing, 1950, McCulloch and Pitts,

1943). Now, ever more complex models are used to model the brain, where the model goes some way beyond capturing brain data to - in some way - simulate and try to reproduce capacities of the human brain.

An absolutely core challenge for studying the brain is to figure out what its relevant parts are, whose causal interactions underpin our cognitive capacities. Parts of the physical brain-as-organ, such as neurons and layers, were reasonably well established quite some time ago, and, while other relevant parts like hormone signaling were discovered more recently, they are now comparatively well understood. Beyond these, though, there is a lot of mystery (Milkowski, 2013, Piccinini, 2020, Barack and Krakauer, 2021). In particular, the brain seems to do quite a lot of relatively holistic processing, and how that is actually performed and where is less well known. It is perhaps not surprising that we use ANNs to try to explore this space of possible processing, to try to get some kind of handle on the brain's extraordinary functionality.

Artificial neural networks (ANNs) are virtual networks composed of simple computational units (nodes) behaving according to their activation functions which loosely stand for real neurons. The connection strengths between them determine how much of an influence they can have on each other (LeCun et al., 2015). Most of the landmark studies nowadays use the performance-optimisation driven approach (Yamins and DiCarlo, 2016), which aims to optimize connection strengths by training ANNs on a particular task. Once an ANN is trained and able to perform on the task, it is used as a model of causal processes in the brain.

We focus on the ANN model family that has so far been typically used in visual neuroscience, deep convolutional neural networks (DCNNs). DCNNs are very successful in “core object recognition” (DiCarlo et al., 2012), which takes images of objects as inputs, and outputs appropriate labels for detected objects in the images (such as “chair” or “motorcycle”). Between the input and output layers, DCNNs solve object recognition by, first, extracting features hierarchically across a number of their layers - detecting simple features such as edges across the whole image, and then more complex features built out of simpler ones, such as curves and more complex shapes. Detection of features is the job of the convolutional nodes that these networks are named after. Successful networks need to learn to recognise those features that enable them to overcome the problem of “nuisance parameters” such as size, shape and position of objects in the visual field. After convolution, DCNNs employ max pooling layers that help with that challenge (Riesenhuber and Poggio, 1999). After hierarchical iterations of convolution and pooling operations, a classifier learns to map extracted features to correct labels that the final layer outputs (Kreiman, 2021, Cao and Yamins, 2021, Buckner, 2019, LeCun et al., 2015).

At least in some cases scientists have seen these modeling efforts as aiming to obtain explanations, besides being helpful with prediction and exploration (Cichy and Kaiser, 2019, Kriegeskorte, 2015). A common thread behind a range of recent papers is the idea that the processing of ANNs can map onto the causal mechanistic structure of the brain, thereby providing mechanistic explanations and understanding. For example, Kietzmann et al. (2019) see DCNNs as being suitable for providing mechanistic explanations of our visual system. And Schrimpf et al. (2020a) discuss ways of benchmarking DCNNs in order to advance the current

state of mechanistic modeling of vision. Taking up the challenge of specifying which explanatory style DCNNs are after, Lindsay (2021) explicitly places them in the mechanistic category, along with Cao and Yamins (2021) and Golan et al. (2023). In addition, Bashivan et al. (2019) extend ways of mapping DCNNs onto brains in a way they take to advance our causal "understanding" of the visual areas. These approaches aim to infer the actual causes of object recognition in the target brain areas by use of ANNs, the problem Termine and Primiero (this volume) call the causal explainability problem.

While we think we're a long way away from mechanistic explanation in the cases we explore, for reasons that will become clear, Bechtel and Richardson's (1993/2010) story about a key mechanism discovery heuristic consisting of decomposition into component parts, plus functional decomposition into component operations, followed by localisation of component operations in component parts, is still a useful framework here. Bechtel (2008) specifically examines the challenges of explaining the brain, showing that, for the brain, possible decompositions into parts, and into operations, are both extremely underdetermined by experimental practices. It is when you manage to localize operations in parts that you reduce the underdetermination problem, and you have some evidence that *both* your decomposition into parts, *and* your decomposition into operations must be heading in the right direction. He illustrates this using studies of memory, explaining that it was only with the discovery of the functionality of the hippocampus that we could become confident that there really was a center for memory encoding in the brain, rather than, for example, visual memories being encoded in the visual system, auditory memories in the auditory system, and so on. We will return to this

idea explicitly in sections 4 and 5, after examining the crucial role of similarity measures in current work.

3 Similarity measures become crucial for claims of causal similarity of ANNs and the brain

The deep learning breakthrough in the engineering field of computer vision (Krizhevsky et al., 2012) has in the last ten years given rise to the novel field of NeuroAI (Momennejad, 2023, Zador et al., 2022). One part of NeuroAI uses ANNs inspired at some level by how the brain works, and successful for engineering goals, and reapplies them to neuroscience to model the brain itself.

Studies using this approach in visual neuroscience pick a particular architecture, often from the DCNN model family, and train it to perform object recognition (e.g. Storrs et al., 2021, Yamins et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014). This is a performance optimisation-driven approach (see more on optimisation in Termine and Primiero, this volume). Many such trained networks show object recognition task performance comparable to humans (Storrs et al., 2021).

Constructing these models and successfully equalling human task performance is already quite an achievement, given just how hard this problem was taken to be (DiCarlo et al., 2012). This success aside, though, it is much less clear what sorts of claims are warranted concerning understanding and explanation of human visual perception based on these models.

The last ten years of work using DCNNs to explain and understand the brain all depends on similarity measures-dependent assertions of similarity, including degrees of similarity, between DCNNs and the brain, with the aim of assessing similarity of their causal mechanisms in object recognition. After training models and assessing their task performance, a particular quantitative similarity measure is adopted to assess how similar the processing is between DCNNs and the brain, when they classify the same images.

This means that similarity measures have so far formed the foundation for doing model selection in this field (Storrs et al., 2021, Schrimpf et al., 2020b). Model architectures in the family of DCNNs vary in depth (i.e. the number of layers, although they all have more than the 3-4 hidden layers characteristic of earlier connectionist networks), but they may also differ in the number of weights and architectural motifs (feedforward or recurrent, residual or plain). While depth was regarded as bringing computational advantages (Buckner, 2019), better task performance does not seem to be strongly related to having more layers (Storrs et al., 2021, p. 2048), or to the number of layers matching the brain, according to at least some similarity measures (Storrs et al., 2021). However, such assessments of similarity of DCNNs and the brain in the process of model selection depend crucially on choice of similarity measures.

In the last decade, the two most widely used measures of similarity between the brain and DCNNs have been representational similarity analysis and linear mapping (for discussion of the field's reliance on these measures, see Kietzmann et al., 2019; Yamins and DiCarlo, 2016).

Representational Similarity Analysis (RSA) is an analysis technique that focuses on responses of a population of neurons in the brain or nodes in a DCNN (Kriegeskorte et al., 2008). For a set of stimuli and a particular brain region or a DCNN layer, RSA quantifies how dissimilar stimuli-elicited patterns are in a pairwise manner. This pairwise information is summarized in the form of a distance matrix, which tells us how *dissimilarly* a set of stimuli are processed by a model, on the one hand, and the brain on the other. In the final step, distance matrices of a model and the brain are compared, and they are judged to be similar to the extent that the model treats the same pairs of stimuli as similar or dissimilar as the brain does.

The choice of similarity measure that is inserted into the RSA framework and used to quantify dissimilarities between stimuli-elicited patterns is an important modeling decision. In this field, that is not usually carefully theoretically justified (Grujičić, 2023; Milkowski, this volume). Various measures are used: correlation (Cichy et al., 2016, Khaligh-Razavi and Kriegeskorte, 2014), cosine (Mehrer et al., 2020), Euclidean and Mahalanobis distances (Storrs et al., 2021), as well as the dot product (Kornblith et al., 2019). The application of these measures can entail different judgements of similarity of compared systems (compare Mehrer et al., 2020 and Kornblith et al., 2019), indicative of the many analysts problem (Silberzahn et al., 2018). But even if this weren't the case, different similarity measures pick out different mechanisms as relevant for object recognition (Ramirez et al., 2014, Grujičić, 2023), thus inviting arbitration between them.

In contrast to RSA, which focuses on the population level of analysis, linear mapping aims to map individual neuronal responses in a brain region based on responses of nodes in a DCNN, when they are exposed to the same stimuli (Cao and Yamins, 2021, Yamins et al., 2014). More specifically, individual neuronal responses are to be predicted based on linear regression from nodes in a DCNN. For a given set of training images, the method finds a set of coefficients which are multiplied with some nodal activations in a DCNN occurring when a particular image is presented, and these quantities are summed together. The result of this operation is taken to be the response of a "synthetic neuron" constructed out of these individual nodal activations, which is then mapped onto the real neuron in the brain (Yamins and DiCarlo, 2016). The coefficients that are learnt are then tested on a new, previously unseen set of images.

The creation of "synthetic neurons" that are mapped onto real neuronal activations raises the question of what kind of information is used to create them, and whether that information is causally relevant to how the model works when it successfully recognises objects. This worry parallels the one arising from the RSA framework, since the variety of similarity measures it uses, which differ regarding the properties of stimuli-elicited patterns they pick out, leaves open the question of whether the comparison between models and the brain is done in relevant ways at all. Thus, the worry is that RSA and LM may compare systems in gerrymandering and irrelevant ways if one aims to conclude something about causal mechanistic similarity of processing of DCNNs and the brain.

In recent years there have been a couple of methods that emerged and which react to this legacy of the field of the last ten years. We discuss two studies which aim to advance mapping DCNNs and the brain in object recognition, by allying linear mapping with broadly experimental practices. This is done with the hope of coming closer to assessments of their causal mechanistic similarity. The studies show that linear mapping can be helpful in establishing interventions in the brain based on the model (Bashivan et al., 2019) and in the model based on the brain (Sexton and Love, 2022). Recall that here the idea of interventions is used in a broad sense matching experimental practices, and not in the more narrowly defined sense of Woodward (2003). These are the only two studies we are aware of that attempt to map DCNNs and brains by introducing an intervention, which is why we choose to discuss them here. In addition, the recently much discussed criticism of RSA and LM-based comparisons by Bowers et al. (2022) omits to acknowledge these intervention-based studies. While Bowers et al. (2022) suggest turning to psychological evidence to arbitrate between models, we explore this other avenue. However, what exactly these intervention-based advances mean for the goals of explanation and understanding is still not clear, as we will examine in the next two sections.

4 Creating artificial images that drive neurons in V4

We will begin by studying Bachivan et al.'s aims. The study of Bashivan et al. (2019) is motivated by overcoming two limitations of previous DCNNs in this field. The first one is that their computations are difficult for humans to comprehend, making it unclear what form of understanding scientists achieve by using them. Second, because DCNNs have only been tested

on visual stimuli similar to the training stimuli, how much these models can generalize is debatable.

In order to address the first concern, Bashivan et al. test if the information contained in a DCNN is already useful to control neural activity, regardless of the model's opacity. They do this by using a DCNN and an image synthesis algorithm to create synthetic images that manage to drive targeted neural sites in V4 in macaques in a desired way. They then use synthetic images in order to address the second concern, by showing that the DCNN can accurately predict brain responses to these novel images.

While taking their study to mitigate the two limitations, Bashivan et al. note that these do not disappear. They conclude that an "important test of understanding is the ability to use knowledge to gain improved control over things of interest in the world, as we have demonstrated; however, we acknowledge that this is not the only possible view, and many other notions of "understanding" remain to be explored to see whether and how these models add value" (Bashivan et al., 2019, p. 7). We will explore this claim.

Bashivan et al. go on to make further claims. Bashivan et al. first build a linear mapping between DCNN nodes and individual neural sites in V4, as many previous studies do. However, they go further by using the obtained linear mapping in order to generate synthetic stimuli that will elicit 'experimenter-desired neural response patterns' (Bashivan et al., 2019, p. 1 of 1) in targeted neural sites in macaque V4. The kinds of activations they choose to elicit are those that drive a

neural site “so strongly as to activate it beyond its typically observed maximal activation level” (p. 1 of 1) in one condition, which was defined by testing its response to a set of naturalistic images. In the second condition they aimed to control one neural site “while simultaneously inactivating the other recorded neurons” (p. 1 of 1).

By focusing on a higher visual area V4 in macaques and probing targeted neural sites in this area, Bashivan et al. primarily focus on decomposing into *parts*. DCNNs have proven useful here, as they “give us new ability to find manifolds of more optimal stimuli for each neural site at a much finer degree of granularity and to discover such stimuli unconstrained by human intuition and the limits of human language" (Bashivan et al., 2019, p. 6). These synthetic images did manage to probe a brain area that shows convoluted response properties that have evaded scientific understanding.

To what extent causally controlling neural responses by presentation of stimuli brings us closer to comparisons of causal mechanisms of ANNs and the brain demands some interpretation. The key issue regards the extent to which it adds to an understanding of the functionality of parts of V4 for object recognition. For Bashivan et al., the very ability to control neural sites in the desired way based on a DCNN suggests, firstly, that linear mapping along with the “knowledge that the models contain is useful for one potential application (neural activity control)” (Bashivan et al., 2019, p. 1 of 11). Secondly, the ability to predict brain responses to novel stimuli generated by the image synthesis algorithm is taken as “a stronger test of functional similarity to the brain

than prior work had shown" (Bashivan et al., 2019, p. 5). Claims like this tell us something about how they are thinking about causality.

It is worth addressing explicitly how Bashivan et al. are thinking about causality. DCNNs have been called the most predictively successful models of the ventral stream (Storrs et al., 2021, Yamins et al., 2014). Bashivan et al. are interested in going beyond the mere predictive success of previous studies, by showing that the information DCNNs capture - that they refer to as "knowledge" that the model contains (Bashivan et al., 2019, p. 1 of 1, and 8 further times in the paper) - is useful for controlling neural responses.

The "knowledge" the model contains is related to latent neural manifolds of object categories that get formed during the process of training (Bashivan et al., p. 6). The study goes beyond prediction based on linear mapping, by controlling responses of neural sites enabled by the formed manifolds in the DCNNs in conjunction with linear mapping. This renders the experiment not merely correlational (p. 2). They clearly value this newly obtained control quite highly. In their opinion, future, more accurate DCNNs will enable better noninvasive neural control (p. 1 of 11), perhaps even useful for therapeutic applications (p. 1 of 1).

Going beyond prediction by way of controlling neurons helps with obtaining some form of understanding of the target mechanism, in their view. Although these models were successful in predicting brain response properties, "their contribution to an understanding of primate visual processing remains controversial." (p. 1 of 1.) And the opacity of DCNNs doesn't help either.

Thus Bashivan et al. put forth the idea of a test of understanding they believe their study passed: "An important test of understanding is the ability to use knowledge to gain improved control over things of interest in the world" (p. 7). It is important to point out again that the knowledge they talk about is the knowledge in the model, and not of the scientists using these models.

They do seem to take the sense of understanding they are talking about to be causal. "We reasoned that successful experimenter control would demonstrate that at least one ANN model can be used to noninvasively control the brain—a practical test of useful, causal "understanding"." (p. 1 of 11) The causal relation holds between stimulus presentation and elicited neural responses, similar to the case of optogenetics (p. 2). The experiment in control, however, does not demonstrate that only some neural sites in v4 can be controlled by the use of a DCNN. When taken to be broadly confirmatory of the whole model, it suggests that the neural manifolds formed along processing stages in the DCNN are on the right track as a model of brain processes.

We will finish by examining what Bashivan et al. have really achieved. In later visual areas, whose parts have been resistant to understanding, Bashivan et al. have gotten hold of a part they can do something with. However, their approach provides only one angle at getting at the parts - from the stimulus. This is continuous with many historical efforts in visual neuroscience, aiming to make the parts intelligible in terms of the representational approach (Bechtel, 2008), analyzing the kinds of stimuli neurons are most responsive to. But the populations of neurons Bashivan et

al. control are very tiny, and the functional relevance of their responses for processes downstream and the task performance is not further traced. By driving neurons above their normally occurring values, one could claim they do not target parts operative in the way they typically work in the system. While the intervention of this type may satisfy an engineer, it may not satisfy a physiologist given that the experimenter-chosen control targets may not reflect a regard for physiologically plausible switches (Craver, 2021) of the values of the variable for processing downstream and the capacity of interest.

Bashivan et al. do not elaborate on how the experiment in control helps us gain a functional understanding of parts, while keeping an eye on the overall capacity of the system within which the part performs its function. This stands in contrast to classic intervention techniques in neuroscience, such as lesions, drug interventions or optogenetics. If we consider Darden's (2006) mechanism discovery heuristics of forwards and backwards-chaining, they have managed to chain backwards from V4 neurons to stimulus, but not managed to chain forwards towards behavior.

This therefore reverses a typical order of first gaining some causal understanding of the system, and then proceeding to control. Things seem to go backwards here. In contrast with the case of modeling in systems biology discussed by MacLeod (this volume), DCNNs help inferring something causally relevant for the target system without their modelers developing a "feel for the model" or a mental representation of its features (MacLeod, this volume). Although they repeat that "knowledge is in the model", in virtue of which control is obtained, not much causal

understanding of the parts is gained, where that is understood as their functional contribution to the capacity to be explained.

5 Switching in brain data for a layer in a functioning DNN

As we did for Bashivan et al., we will begin by examining Sexton and Love's (2022) aims. They begin with a limitation of the metrics previously used to compare DCNNs with the brain, which assume that correlation is a good measure of correspondence. Both representational similarity analysis (RSA) and linear mapping (LM) can only assess this shared variance. However, shared variance does not imply functional similarity between systems, because the variance that is shared may not be relevant for the behavior of interest. This echoes our worry above that RSA and LM may compare systems in gerrymandered and irrelevant ways for object recognition.

To address this shortcoming, Sexton and Love introduce a stronger test of brain-likeness which relies on directly interfacing model and brain activations. If a DCNN successfully classifies objects after a substitution of its layer activations with brain data, then we know their similarity is task-relevant similarity, which is better evidence for their functional similarity.

The focus on task-relevant similarity yielded a different pattern of DCNN-brain correspondences in comparison to the ones obtained by RSA and LM only. Previously, DCNNs and the visual areas were found to correspond in a hierarchical manner with early model layers most similar to

early brain areas, and deeper layers to later brain areas. However, Sexton and Love found that all ventral stream areas best corresponded to the late DCNN layers. Their further analyses favored the conclusion that "long-range recurrence between higher-level brain regions, such as IT, influenced activity in lower-level areas like V4" (p. 4-5). This challenges traditional hierarchical accounts, in favor of non-hierarchical, recurrent processing even in core object recognition.

Sexton and Love then go on to make further claims. We now need to dig a bit deeper into how Sexton and Love go beyond the previous reliance on metrics, particularly the 'shared variance' that they are so critical of. In section 4, we discussed parts first, because Bashivan et al. (2019) focus on various parts, particularly the novel images they generated, and the targeted neural sites. Although functionality was a concern, they made few direct claims about it, requiring interpretation. Sexton and Love approach their work in the opposite direction, introducing a metric that focuses on task-relevant functionality as a way to make ANN and brain comparisons more causally relevant. They use the term "task-relevant" repeatedly (20 times), writing about task-relevant activity (7), variance (6), information (6), analysis (1) and signals (1).

They want to test whether a DCNN layer and brain region match in functionality, showing that they do by swapping brain data into the DCNN without affecting its behavior, i.e. the object classification. Metrics remain crucial, with linear transformation used to get brain data into the model. Nevertheless, it is important to them that previous studies did not achieve their connection to behavior. Despite this focus, their work can still be seen as identifying parts, in the wider sense allowed by Bechtel and Richardson's heuristics where, before localisation, we

decompose into both component parts and component *operations*. Operations are often neglected, but we can understand Sexton and Love as making them central when they focus on 'task-relevance'.

So Sexton and Love are really comparing *functionality* in the brain and the model. In a sense, they are taking brain data representing the activity of a particular area of the brain, and seeing whether the DCNN can support the same activity - the same component operation - with the behavior, i.e. the object recognition, unchanged: "replacing model activity with brain activity should successfully drive the DCNN's object recognition decision." (Sexton and Love, 2022 p. 1-2.) This is as direct a test as can currently be done of whether the brain operation and the DCNN operation are the same. Indeed, it echoes, as far as is possible, standard methodologies in life sciences of studying the operations of a suspected part of a mechanism by extracting it from its usual location and studying its activity in isolation, albeit in a substrate that can support it (Güttinger, 2013). Of course, in this case, what is extracted is not a component part, but a component *operation*.

Finally, they do seem to be attempting, at some level, some localisation. Writing, "For a chosen model layer and brain region," (Sexton and Love, 2022, p. 1) indicates that their comparison is of operations already localized, both in the DCNN and in the brain. This is confirmed, "How well the DCNN performs when directly interfaced ... with the brain provides a strong test of how well the interfaced brain region corresponds to that layer of the DCNN" (Sexton and Love, 2022, p. 1). Their original conclusion seems to be about component operations and their localisation, too,

specifically about where recurrent processing happens. They summarize: “We found that all brain regions, from the earliest to the latest of visual areas along the ventral stream, best corresponded to the later model layers. These results indicate that neural recordings in all regions contain higher-level information about object category even when most variance in a region is attributable to lower-level stimulus properties.” (Sexton and Love, 2022, p. 4.)

As for Bashivan et al., we will also make explicit how Sexton and Love are thinking about causality. Sexton and Love are cautious, making few claims directly about causation or understanding. They begin, though, in a similar place to Bashivan et al., appreciating the predictive success of previous DCNNs. They first establish the linear mapping between model and brain, for the DCNN they use, but they go beyond it by introducing an intervention.

Like Bashivan et al., Sexton and Love are attempting to use experimental work to enhance previous results, but they do not aim for the kind of control that Bashivan et al. value so highly. While Bashivan et al. use DCNNs to synthesize novel images to intervene in real brains, Sexton and Love work in the opposite direction, swapping in brain data to the DCNN. And they aim not to change something but to maintain something - i.e. the DCNN’s task performance. This is still a kind of control, but much more broadly conceived, perhaps as the control of a component operation. Notice, though, that Bashivan et al. work from stimulus image to neural activity, where Sexton and Love intervene in the DCNN, targeting from DCNN layer to unaltered object classification.

Sexton and Love talk very little about causal understanding or even plain understanding, seeming happy to think of their achievement more narrowly as something like better evidencing the functionality of brain regions. Nevertheless, it seems to us that they have offered some advance in understanding.

While this is not yet a completely convincing understanding of a localized operation (particularly as their results are in tension with those of Bashivan et al., as we will return to shortly), Bechtel and Richardson's heuristic framework is still helpful. Using it, we can see that they attempt to do what's very standard in other forms of mechanism discovery, which is substituting parts in and out of systems, to study what they do in different contexts. For Sexton and Love, this is a way of studying component operations and their localisation, taking brain data and effectively simulating the brain operation in the DCNN. What is distinctive here is the learning how to map brain data into a DCNN, and switching that transformed operation into an entirely artificial system, what they call 'interfaced' into the DCNN.

Sexton and Love write "If a brain region corresponds functionally to a model layer, then the brain activity substituted for the model activity at that layer should drive the model to the same output as when an image stimulus is presented" (Sexton and Love, 2022, p. 4). This of course means the mapping of brain data into DCNN is crucial, and they know that this all hangs on the translation metric. Nevertheless, it is an achievement, particularly to get a still-functioning system in the sense that it is still performing the overall task. So we agree with Sexton and Love when they write: "By minding the distinction between shared and task-relevant variance (i.e.,

activity that can drive the computation), the role that the brain regions play within the overall computation may more readily come into focus." (Sexton and Love, 2022, p. 5.) They have made a step on a way to a mechanistic explanation, a causal explanation of how our brains manage to process visual information, in an area where this is especially challenging.

We will finish, again, by considering what Sexton and Love have really achieved. Ultimately, while we do not think that Sexton and Love have achieved mechanistic explanation, they have really succeeded in probing the functionality of both brains and DCNNs. We think they are right to give high priority to task-relevance, attempting forward-chaining from neural activations to behavior (Darden, 2006). In managing to move directly between the brain and DCNN they, like Bashivan et al., do move significantly beyond relying exclusively on similarity metrics to analyze causal mechanisms of the two systems.

Of course, what Sexton and Love mean by behavior is still fairly thin, just the DCNN outputting a label, allowing them to compare its success before and after brain data is swapped in, and of course this depends entirely on the successful translation of brain data into the DCNN (given that task performance drops to around chance level in the case of untrained, randomly initialized networks (Yamins et al., 2014, Fig. 1A), so one can't argue that the architecture alone rather than the weights is the main contributor to the effect). Nevertheless, this does go beyond the previous work in a remarkable way, in allowing us to study the functionality of parts of both DCNN and brain.

6 Conclusion

We have to assess the work here against the backdrop of understanding the scale of the challenges it faces. First, let us return to a point we began section 1 with: Chang (2004) is an excellent lesson in how complex and messy the history is of acquiring what we now treat as simple and reliable measurement tools, such as thermometers to measure temperature. Secondly, recall that in section 2 we mentioned Bechtel's (2008) point that possible decompositions into parts and operations of the brain are *severely* underdetermined by experimental practices. Bashivan et al. and Sexton and Love face both of these problems.

Against this backdrop, our criticism is not that this work depends on metrics. That should be expected, and will continue to be the case. The questions are what metrics, how are they used, and what techniques are used to reduce a sole dependence on metrics, when different metrics yield different results and pick out different mechanisms. In this, we have shown that both Bashivan et al. and Sexton and Love have made significant achievements in more directly relating causal workings of the brain and DCNNs.

Nor is it a devastating criticism that mechanistic explanation has not - or not yet - been achieved. Instead, we have argued that the two key papers probe different parts of the system: Bashivan et al. the part from stimulus to neuron activation; Sexton and Love trying to see whether brain area activation data can function in the place of a DCNN layer. For us, these both move beyond

previous work that is fundamentally limited in relying solely on using similarity metrics to compare brains and DCNNs.

The scale of the underdetermination means we need constraints, whether they come from Bechtel and Richardson's localisation, or from somewhere else. A major impediment to this is that our computational methods are now so powerful, and abundant, that they do not always help introduce constraints (see also Miłkowski, this volume). There are many similarity metrics to choose from, they will all produce some result or other, and these often conflict (cf. Mehrer et al. (2020) and Kornblith et al. (2019)).

One alternative to this is to do experimental psychological work to assess their similarity with the brain mechanism (Bowers et al., 2022). Another seems to be to do the kind of innovative experimental work that attempts to cross from DCNNs to brain and back again that we discuss here. These attempts to replicate functionality in the brain using something derived from DCNNs, and functionality in DCNNs using something derived from the brain, introduce constraints that can reduce underdetermination. While we cannot literally directly couple brain and DCNN, this will rely on a mapping of brain data, or on algorithms to extract 'knowledge' from DCNNs, but these are still good things to do.

One sticking point, however, is that Bashivan et al. and Sexton and Love don't agree. We have argued that they are both probing both parts and operations, albeit that Bashivan et al. move from parts to functionality, while Sexton and Love move in the opposite direction. They are also both

interested in localizing component operations in component parts. Both papers use linear mapping alongside some kind of intervention which attempts to cross from a DCNN to brain, in a way which should introduce useful constraints on the localisation. But recalling Bechtel's (2008) point about the importance of localisation of operations in parts, they do not come to the same conclusions!

Bashivan et al. map the operations of the model onto brain parts in a hierarchical way - they use model V4 area in order to control V4 neurons. This is continuous with the previous findings in the field (Yamins et al., 2014) - the mapping of layers onto brain areas enfolds them in an early-to-early and late-to-late fashion. In contrast, the mapping of operations of the model onto brain parts done by Sexton and Love is non-hierarchical. In contrast with the previous findings, all brain areas map successfully onto later model layers, suggestive of task-relevant information being present already in early areas. Ultimately they arrive at different mappings of model parts to brain parts.

Of course models can have different goals, and in some cases inconsistent models would not be a problem. However, in this work in visual neuroscience, many scientists are trying to understand the mechanism of object recognition, and for that goal inconsistent models will not do. This shows, then, that there is a lot more work needing done, that these are only the initial attempts to go beyond comparisons relying exclusively on similarity metrics.

This problem would be less pressing if DCNNs were made in a more hypothesis-driven way, with location constraints built in. So initially, location constraints were not imposed to build DCNNs because getting task performance equivalent to human performance worked without them. But now that has been achieved, if we wish to build DCNNs specifically to understand the human brain, we might need to reintroduce constraints. One of the reasons mapping model operations onto brain parts in various ways is possible is because DCNNs are not built based on an explicit hypothesis about which brain area particular layers should map onto. In the performance optimisation-driven approach a DCNN is optimized to be able to perform well in the task, so the solution it arrives at, layer per layer, is not meant to be mapped in advance onto particular brain areas. The mapping one obtains between model layers and brain areas depends upon the adopted metric. If a DCNN were constrained based on an explicit hypothesis which brain area it should map onto, the problem of underdetermination would be seriously diminished.

Until that becomes possible, and more work in this vein is completed, mechanistic explanation of the human visual system remains distant. This, however, is not the only way that understanding can be advanced. According to the later iterations of the contextual theory of understanding, scientists understand a theory if they can use it to build models (de Regt, 2017). Illari (2019) uses stellar structure models in astrophysics to extend the contextual theory of understanding to a case where our understanding of stars and supernovae largely lies in clusters of models, rather than in a separate theory of stars, and acknowledges that scientists have multiple uses for those models. Following this example would support the view that scientists broadening their approach beyond

looking for shared variance and becoming able to use DCNNs for further ends does show increased understanding beyond previous work relying solely on measures of shared variance.

Ultimately, we have illustrated that the quest to improve causal understanding of causal mechanisms of object recognition, and causal inference in the sense of trying to discover those mechanisms, is a richly complex task, mediated here by both technologies in DCNNs, advanced data processing in the construction of similarity metrics, and innovative experimental techniques trying to probe the space of components - both component parts and component operations - of the brain. As Bechtel (2008) points out, the component parts and operations of the brain are particularly severely underdetermined. This kind of work manages to introduce *some* kinds of useful constraints. That aside, there doesn't seem to be any standard relationship between inference, experiment and understanding, here, rather there is a set of highly customized experimental practices.

References

Barack, D. L. & Krakauer, J. W. 2021. Two views on the cognitive brain. *Nature Reviews Neuroscience,* 22**,** 359-371.

Bashivan, P., Kar, K. & Dicarlo, J. J. 2019. Neural population control via deep image synthesis. *Science,* 364**,** eaav9436.

Bechtel, W. 2008. Mental mechanisms: Philosophical perspectives on cognitive neuroscience, Psychology Press.

Bechtel, W. & Richardson, R. C. 1993/2010. Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research, Princeton.

Bickle, J., Craver, C. & Barwich, A. S. 2021. The Tools of Neuroscience Experiment: Philosophical and Scientific Perspectives, Routledge.

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

Buckner, C. 2019. Deep learning: A philosophical introduction. *Philosophy Compass,* 14**,** e12625.

Cao, R. & Yamins, D. 2021. Explanatory models in neuroscience: Part 1--taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

Chang, H. 2004. Inventing Temperature: Measurement and Scientific Progress, OUP.

Cichy, R. M. & Kaiser, D. 2019. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences,* 23**,** 305-317.

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**,** 27755.

Craver, C. 2021. Towards an epistemology of intervention: Optogenetics and maker's knowledge. The Tools of Neuroscience Experiment: Philosophical and Scientific Perspectives. Routledge.

Darden, L. 2006. Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution, Cambridge University Press.

De Regt, H. W. 2017. Understanding scientific understanding, Oxford University Press.

Dicarlo, James j., Zoccolan, D. & Rust, Nicole c. 2012. How Does the Brain Solve Visual Object Recognition? *Neuron,* 73**,** 415-434.

Golan, T., Taylor, J., Schütt, H., Peters, B., Sommers, R. P., Seeliger, K., Doerig, A., Linton, P., Konkle, T. & Van Gerven, M. 2023. Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. https://doi.org/10.31234/osf.io/tr7gx

Grujičić, B. 2023. Deep convolutional neural networks are not mechanistic explanations of object recognition. http://philsci-archive.pitt.edu/22629/
Güttinger, S. 2013. Creating parts that allow for rational design: Synthetic biology and the problem of context-sensitivity. *Studies in History and Philosophy of Biological and Biomedical Sciences,* 44**,** 199-207.
Illari, P. 2019. Mechanisms, Models and Laws in Understanding Supernovae. Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie, 50**,** 63-84.
Illari, P. & Russo, F. 2014. Causality: Philosophical theory meets scientific practice, Oxford University Press.
Khaligh-Razavi, S.-M. & Kriegeskorte, N. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology,* 10**,** e1003915.
Kietzmann, T. C., Mcclure, P. & Kriegeskorte, N. 2019. Deep Neural Networks in Computational Neuroscience. Oxford University Press.
Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. International Conference on Machine Learning, 2019. 3519-3529.
Kreiman, G. 2021. Biological and Computer Vision, Cambridge University Press.
Kriegeskorte, N. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science,* 1**,** 417-446.
Kriegeskorte, N., Mur, M. & Bandettini, P. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
Krizhevsky, A., Sutskever, I. & Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems,* 25.
Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *Nature,* 521**,** 436-444.
Lindsay, G. W. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience,* 33**,** 2017-2031.
MacLeod, M. 2024. Seeing further: the role of modelers and simulation in causal inference. The Routledge Handbook of Causality and Causal Methods. Routledge.
Mcculloch, W. S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics,* 5**,** 115-133.
Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. 2020. Individual differences among deep neural network models. *Nature Communications,* 11**,** 5725.
Milkowski, M. 2013. Explaining the computational mind, MIT Press.
Milkowski, M. 2024. Comparing Prediction and Explanation in Computational Models: Theoretical Neuroscience vs. Language Technology. The Routledge Handbook of Causality and Causal Methods. Routledge.
Momennejad, I. 2023. A rubric for human-like agents and NeuroAI. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378**,** 20210446.

Piccinini, G. 2020. Neurocognitive mechanisms: Explaining biological cognition, Oxford University Press.
Ramírez, F. M., Cichy, R. M., Allefeld, C. & Haynes, J.-D. 2014. The Neural Code for Face Orientation in the Human Fusiform Face Area. *The Journal of Neuroscience,* 34**,** 12155.
Riesenhuber, M. & Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neuroscience,* 2**,** 1019-1025.
Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R. & Dicarlo, J. J. 2020a. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron,* 108**,** 413-423.
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K. & Dicarlo, J. J. 2020b. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv***,** 407007.
Sexton, N. J. & Love, B. C. 2022. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances,* 8**,** eabm2219.
Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C. & Bonnier, E. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science,* 1**,** 337-356.
Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. 2021. Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience***,** 1-21.
Tal, E. 2020. Measurement in science. The Stanford Encyclopedia of Philosophy.
Tal, E. 2013. Old and new problems in philosophy of measurement. *Philosophy Compass,* 8**,** 1159-1173.
Termine, A., Primiero, G. 2024. Causality problems in Machine Learning Systems. The Routledge Handbook of Causality and Causal Methods. Routledge.
Turing, A. M. 1950. Computing machinery and intelligence. *Mind,* 59**,** 433-460.
Woodward, J. 2003. Making things happen: A theory of causal explanation, Oxford University Press.
Yamins, D. L. K. & Dicarlo, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience,* 19**,** 356-365.
Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. & Dicarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences,* 111**,** 8619-8624.

Zador, A., Richards, B., Ölveczky, B., Escola, S., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A. & Clopath, C. 2022. Toward next-generation artificial intelligence: catalyzing the NeuroAI revolution. *arXiv preprint arXiv:2210.08340*.