On the importance of severely testing deep learning models of cognition

Jeffrey S. Bowers[1], Gaurav Malhotra[1], Federico Adolfi[1,2], Marin Dujmović[1], Milton Montero[1], Valerio Biscione[1], Guillermo Puebla[3], John H. Hummel[4], and Rachel F. Heaton[4]

[1]School of Psychological Sciences, University of Bristol, Bristol, UK

[2]Ernst Strüngmann Institute for Neuroscience in Coop. with Max Planck Society, Frankfurt, Germany

[3]National Center for Artificial Intelligence, Vicuña Mackenna 4860, Macul, Chile

[4]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, USA

Abstract

Researchers studying the correspondences between Deep Neural Networks (DNNs) and humans often give little consideration to severe testing when drawing conclusions from empirical findings, and this is impeding progress in building better models of minds. We first detail what we mean by severe testing and highlight how this is especially important when working with opaque models with many free parameters that may solve a given task in multiple different ways. Second, we provide multiple examples of researchers making strong claims regarding DNN-human similarities without engaging in severe testing of their hypotheses. Third, we consider why severe testing is undervalued. We provide evidence that part of the fault lies with the review process. There is now a widespread appreciation in many areas of science that a bias for publishing positive results (among other practices) is leading to a credibility crisis, but there seems less awareness of the problem here.

On the importance of severely testing deep learning models of cognition

Introduction

Modelling in neuroscience has increasingly involved deep neural networks. But this line of research, sometimes called “neuroconnectionism” (Doerig et al., 2022) or “neuroAI” (Zador et al., 2023), suffers from many conceptual and methodological problems that contribute to unwarranted conclusions and claims regarding brain representations and processes (see Bowers et al., 2022, for an extended community discussion). Problems include logical fallacies (Guest & Martin, 2023), overclaiming (e.g., Rawski & Baumont, 2022), unchecked degrees of freedom (e.g., Schaeffer, Khona, & Fiete, 2022), naive empiricism and inadequate theorizing (cf. van Rooij & Baggio, 2021), mismatch between measurements and interpretations (e.g., Dujmović, Bowers, Adolfi, & Malhotra, 2023). In this article we focus on another problem that has has not received enough attention, namely, the *lack of appropriate testing of empirical claims*. As detailed below, it is becoming increasingly evident that many prominent claims regarding DNN-human similarities do not stand up to closer scrutiny, and in order to address this problem, we argue that the philosophy of severe testing is needed.

The unique challenges of research comparing DNNs to humans

All empirical sciences rely on carrying out experiments to test hypotheses and evaluate models of natural systems, such as brains. But there are some unique features of DNNs as models of brains that make empirical testing of claims especially challenging.

Consider DNNs as models of human vision. Compared to all previous models, DNNs have the property that they can recognize naturalistic images of objects at a similar rate to humans (sometimes better) on some image datasets, such as ImageNet (Deng et al., 2009). This has led researchers to hypothesize that DNNs may also identify objects in a similar way to humans. And indeed, there is now a large literature of empirical results comparing DNNs to humans, and many findings have been taken to suggest that models do indeed learn similar representatiwons to brains. For example, the observation that activation patterns of units in DNNs are better at predicting neuron activations in visual cortex compared to other models is often used to argue that DNNs are the “current best” models of biological vision.

However, there are reasons to be skeptical regarding these claims. The first reason to be cautious is the opaqueness and expressivity of DNNs. In contrast to other types of cognitive models that consist of a handful of parameters with clear conceptual meaning, deep learning models consist of millions of parameters which are by and large uninterpretable. In fact, more recent systems—such as Vision Transformers and Large Language Models—have several billion parameters. This gives these systems high expressivity and multiple realizability. That is, there are many possible ways in which a deep learning system can learn to map a set of inputs to their outputs.

This high expressivity coupled with the opaqueness inherent in the large number of parameters makes it challenging to understand how a given input is transformed (mapped) to an output. In the absence of this understanding, it becomes difficult to provide in-principle explanations for how a model accounts for a given psychological phenomenon, and whether the model is using similar mechanisms to the visual system. For example, there are recent demonstrations that some DNNs rely on shape rather than texture when classifying objects (Hermann, Chen, & Kornblith, 2020), similar to humans. But when a DNN learns a shape-bias, is it because shape features are more predictive in the training dataset, or because they are easier to extract from a typical stimulus or because of an architectural property of the system? The mere fact that a DNN shows a shape-bias does not provide much evidence that the DNN identifies objects like humans as there are many different ways this outcome may have been realized, many of which will be unrelated to how or why a human shows a shape bias.

The second reason to be skeptical is that there is very little reason, *a priori*, to believe that DNNs will be good models of human cognition. Some researchers interested in drawing parallels between the two systems emphasize the architectural or mechanistic overlaps between DNNs and the primate brain—e.g., units in DNNs are often convolutional, just like simple cells in the primary visual cortex, that learning in both systems occurs in the weights (synapses) between neurons (units) that are hierarchically organized to encode more and more complex features. But beyond these basic similarities, DNNs and brains are different in countless ways, including the fact that (1) neurons in the cortex vary dramatically in their morphology whereas units in DNNs tend to be the same apart from their connection weights and biases, and (2) neurons fire in spike trains where the timing of action potentials matter greatly whereas there is no representation of time in feed-forward or recurrent DNNs other than processing steps. Similarly, current DNNs learn based on algorithms and loss-functions (back-propagation, ReLU units, dropout, batch-normalization) that also have very little psychological / biological grounding. This no doubt relates to the fact that current DNNs need much more supervised training to support a task compared to humans. In combination with the high expressivity of DNNs, there is no reason to assume that DNNs converge onto the same human solution when trained to perform a task such as object recognition.

To further complicate matters, claims regarding DNN-human correspondences frequently rely on the concept of *emergence* — that is, training a network to do one task (e.g., object-recognition) leads to a known psychological phenomenon (e.g., shape-bias). It is important to note how this reliance on emergence contrasts with typical models in psychology and neuroscience, where models embody specific hypotheses and it is comparatively clearer to the researcher exactly the predictions the model will make. In contrast, researchers comparing DNNs to humans frequently do not understand the mechanism through which an observed phenomenon emerges. Due to this opaqueness of the models, researchers rely heavily on testing the models empirically. But if these empirical tests are not carried out rigorously, they may lead to incorrect inferences at several stages in this research pipeline. First of all, it is possible that DNNs perform a task (e.g., object-recognition) like humans on some dataset, but their performance is entirely unlike humans on other datasets (e.g., when noise is added to images; Geirhos et al., 2018). Secondly, it is possible that the hypothesised emerged phenomenon (e.g., shape-bias) only emerges under some very limited conditions. Finally, it is possible that even though a hypothesised phenomenon emerges, it differs qualitatively or quantitatively from the phenomemon of interest in humans. For example, it is possible that both DNNs and humans show shape-bias, but the properties of this shape-bias are qualitatively (Malhotra, Dujmović, & Bowers, 2022; Malhotra, Dujmović, Hummel, & Bowers, 2023) and quantitatively (Geirhos et al., 2019) different between the two systems.

The above considerations emphasize the importance of carrying out rigorous tests that avoid the incorrect inferences listed above. A proper grasp of what conditions make empirical tests appropriate for drawing these conclusions is crucial here. In this article, we argue that this is precisely where current approaches are falling substantially short of the minimum requirements. We will illustrate these problems with a series of examples.

Why is there so little severe testing in this domain? We argue that part of the problem lies with the peer-review system that incentivizes researchers to carry out research designed to highlight DNN-human similarities and minimize differences. We substantiate this claim with examples that illustrate how reviewers and editors undervalue the contribution of studies that rigorously test hypotheses related to deep learning approaches to cognition. But before we do this, we begin by describing what counts as a rigorous test. In particular, we describe the notion of *severe testing* (Mayo, 2018) and argue that following the principles of severe testing is likely to steer empirical deep learning approaches to brain and cognitive science onto a more constructive direction.

What counts as severe testing

The notion of *severe testing* (Mayo, 2018) allows us to conceptually[1] sort out what it means for a claim (e.g., that a certain algorithmic model uses the same features as humans to categorize images) to be supported by evidence (e.g., the outcome of an experiment presenting images to algorithmic implementations and humans). Contrary to the a model comparison approach that is popular in deep learning applications to cognitive/neural modeling (see, for example, Schrimpf et al., 2018), it will be argued that the mere advantage of one model over the other in predicting domain-relevant data is wholly insufficient even as the weakest evidentiary standard.

An entry point to the severe testing idea is through the *weak severity requirement*. Put simply, it asks the researcher to reject the possibility that there is evidence for a claim if nothing has been done to uncover ways in which the claim might be false. For instance, if certain data agree with a certain claim but the test method is practically guaranteed to find such agreement, and had little or no capability of finding flaws with the claim in the case they exist, then

[1] For our purposes, it will be sufficient to consider the conceptual scaffolding around the severe testing idea independent of its ramifications in the philosophy of statistics where it originates. Hence, we make no claims regarding, for example, Frequentist vs Bayesian statistical approaches to data analysis. Our discussion is concerned with rigorous testing of claims regardless of what approach to data analysis is favored.

according to the severity requirement we have no evidence at hand. This is the basic principle that disabuses researchers of the notion that empirical tests, never mind their inadequacies, provide confirmation of a claim at least to a certain extent.

This first aspect of severe testing warns us not to mistake the outcomes of inadequate tests for evidence. The second aspect of severe testing tackles what it means to have evidentiary support for a claim. It says that we only have evidence for a particular claim to the extent that the latter survives a stringent scrutiny. If the claim passes a test whose procedure was highly capable of finding departures from the claim where none or few are found, then we have evidence at hand. That is, for a certain empirical test outcome to warrant a claim, it is required not just that the claim agrees with the outcome. It is crucially required that it be very unlikely the claim would have passed the test if it were false.

Many questions arise as we attempt to unfold what severity requirements mean in practice. How many tests are enough? How stringent should they be? What are the relevant dimensions of stringency? How many flaws are too many? We acknowledge from the outset that these are difficult questions that research communities will only find partial answers to, tailored to specific domains. At the same time, it is important to note that current testing does not even come close to any reasonable severity requirement (cf. Bowers et al., 2022, and the following sections). Therefore, it is important to encourage the community to reflect on the notions of severe testing explained here and to adopt a more self-critical approach to empirical claims.

The severity requirements stated above imply that to have any evidence at all, even a mere indication, we must have more than just a boost in data predictivity under some condition relative to others (e.g., architecture change, training dataset modification, etc.). We require instead a minimum threshold of severity to be met by our tests. In the next section, we will present some common patterns found across this area of research which illustrate how a lack of severe testing manifests itself.

How lack of severe testing plays out: some illustrative examples

To illustrate how the practice of severe testing has played out in recent research, we focus two important lines of research used to support the conclusion that DNNs and humans share similar visual representations, but briefly consider additional examples in the domain of vision, memory, and language as well.

First, multiple studies have compared the patterns of unit activations in DNNs to neuron activations in visual cortex (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins et al., 2014). There are multiple measures that have been used to make these comparisons and we focus on two: representational similarity analysis (RSA) and fitting regression models to predict neural activity from internal activations of DNNs. To employ RSA, one first has to collect neural recordings (e.g., fMRI, EEG, single cell recordings in case of monkeys) and internal activations from DNNs in response to a set of stimuli. Then, pair-wise distances for each pair of stimuli are computed (e.g., 1-Pearson's r between activation vectors for a pair of images) both for humans and DNNs. This results in two representational dissimilarity matrices (RDMs), one for each system being compared. The RDM represents the relative distances between representations of objects in the dataset for a given system (see Figure 1). Finally, the correspondence between RDMs is assessed, usually as a rank-order correlation between them.

The second measure uses DNN activations as predictors for neural activity in a regression model and measures the amount of explained variance (Schrimpf et al., 2018; Yamins et al., 2014). While these two methods are different, the claim that was made early on, based on both methods, was that early layers of DNNs correspond better to neural activity in early areas of vision (e.g., V1) while deeper layers correspond better to later visual processing (e.g., IT). For example, Figure 2 shows results from Khaligh-Razavi and Kriegeskorte (2014), where this claim of hierarchical correspondence was based on RSA. Another early observation was that better performance in classification was associated with better neural predictivity Yamins et al. (2014). The general assumption of this work has been that the better the brain prediction the better the DNN-human correspondence. For example, Brain Score website Schrimpf et al. (2018) includes a leader-board that ranks models in terms of their correspondence to "core object recognition" based on their overall regression predictivity of a number of brain datasets as well as their performance on a number of behavioural benchmarks.

A number of more recent brain-predictivity studies have been carried out that investigate properties of models (architectures, learning algorithms, loss functions, etc.) and training datasets that impact on correspondence between primate visual representations and DNNs as measured by

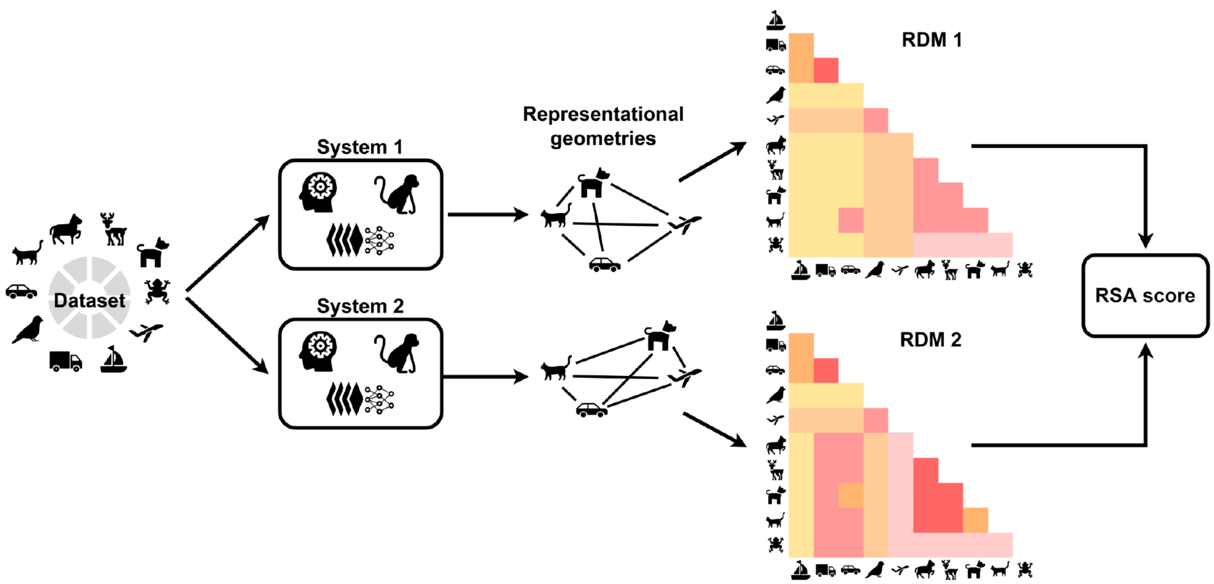

Figure 1

RSA calculation. *Stimuli from a set of categories (or conditions) are used as inputs to two different systems (for example, a human brain and a primate brain). Activity from regions of interest is recorded for each stimulus. Pair-wise distances in activity patterns are calculated to get the representational geometry of each system. This representational geometry is expressed as a representational dissimilarity matrix (RDM) for each system. Finally, an RSA score is determined by computing the correlation between the two RDMs. It is up to the resercher to make a number of choices during this process including the choice of distance measure (e.g., 1-Pearson's r, Euclidean distance etc.) and a measure for comparing RDMs (e.g., Pearson's r, Spearman's ρ, Kendall's τ, etc.). Figure adapted from Dujmović et al. (2023)*

these metrics. For example, Mehrer, Spoerer, Jones, Kriegeskorte, and Kietzmann (2021) show that this correspondence can be increased by training DNNs on a more ecological image dataset. In another study, Zhuang et al. (2021) showed that comparable (though not quite as high) correspondence can also be shown by some self-supervised models.

It should be noted, however, that few studies have attempt to falsify or conduct a severe test on the hypothesis that DNNs and primary visual cortex learn similar representations (but see paper on controversial stimuli from Golan, Raju, and Kriegeskorte (2020)). Ignoring for a moment that claims regarding "core object recognition" are far too expansive and unconstrained

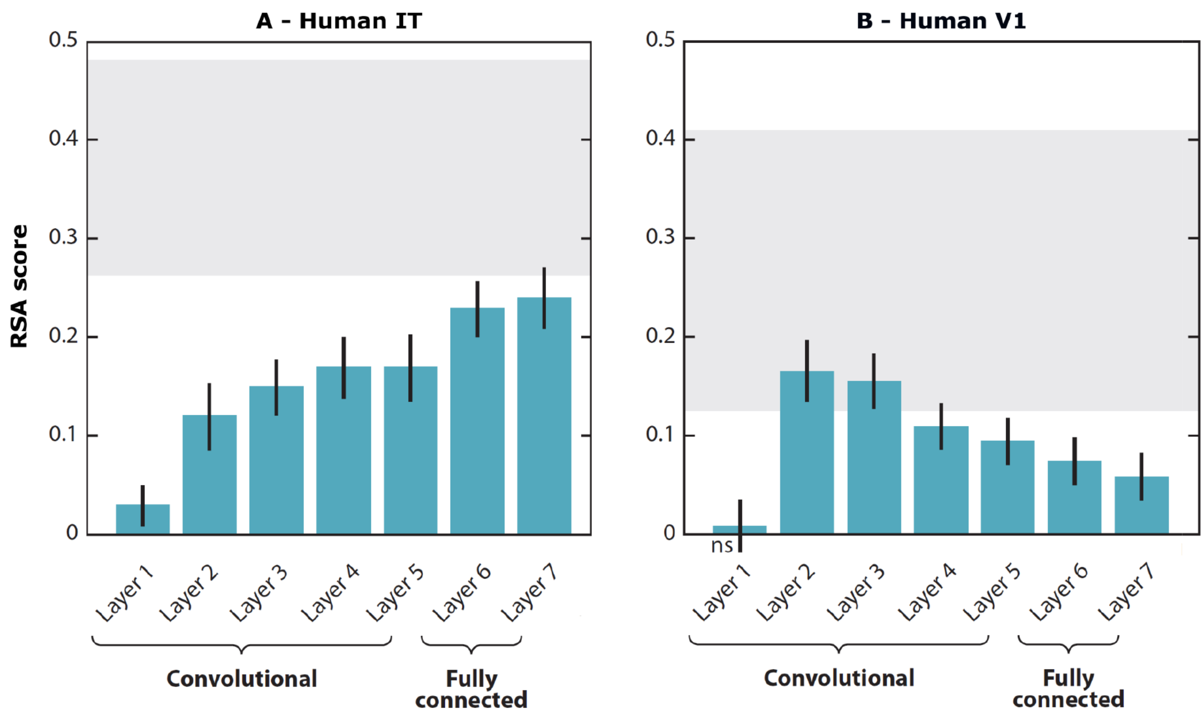

Figure 2

RSA scores of AlexNet layers with neural activity from human IT (A) and V1 (B).

RSA scores between AlexNet layers and human neural fMRI patterns were computed as the Kendall τ between RDMs. The shaded region represents the estimated noise ceiling (expected human to human RSA scores). The figure was adapted from Khaligh-Razavi and Kriegeskorte (2014).

given the nature of the predictivity measures, the overarching goal has been to *increase* the alignment between models and neural representations as measured through prediction scores. In fact, many of these studies rely on a small number of neuro-imaging datasets that have presented a curated set of objects and categories to a small number of primates and humans. For example, the entire suite of 5 IT benchmarks in Brain Score comes from neural data of 5 macaques observing very similar stimuli. If, instead, the goal was to do a severe test, studies would have varied properties of datasets in order to verify whether central observations—such as a hierarchical correspondence between activations of DNNs and visual cortex—bear out. In a recent study, Xu and Vaziri-Pashkam (2021) carried out such a controlled test. They observed that the claim of a hierarchical correspondence between the ventral visual cortex and layers of DNN did not hold up when properties of the input stimuli were changed (see Figure 3), directly undermining previous claims. Similarly, when Sexton and Love (2022) used a different metric to measure correspondence—instead of RSA, their method substituted the activity of a layer with an activity of a brain region—they also observed no hierarchical correspondence between DNN and brain activity. More worryingly, Dujmović et al. (2023) show that previous observations of correlations using RSA could plausibly be due to confounds present in datasets, rather than a mechanistic similarity between the two systems.

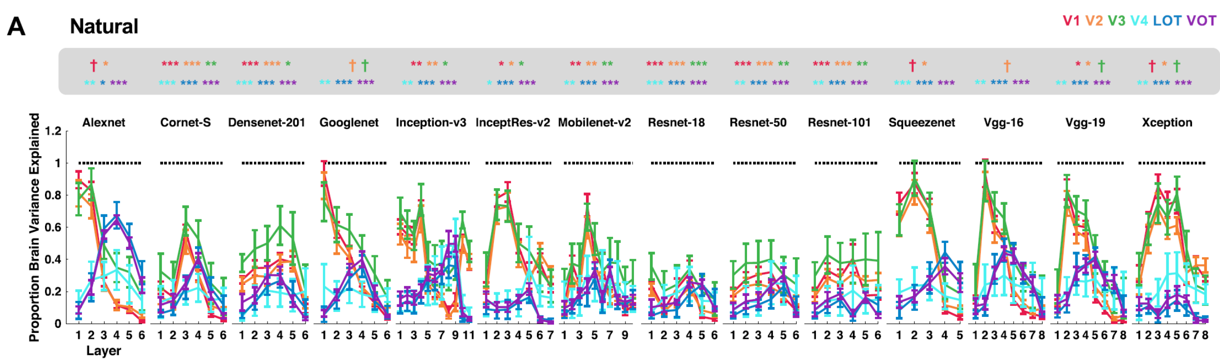

Figure 3

DNN to human correspondence as a function of network layer and brain region from Xu and Vaziri-Pashkam (2021). *Contrary to the claim that early layers of DNNs correspond better to early areas of visual processing (e.g., V1) compared to later layers which correspond better to later areas (e.g., ventral occipito-temporal - VOT), results from Xu and Vaziri-Pashkam (2021) show that there is no such hierarchical correspondence.*

In the second line of research there has been focus on a more specific claim regarding visual DNN-human similarities, namely, whether DNNs and humans share a similar shape *shape-bias*. It has been long known to both vision scientists (Biederman & Ju, 1988; Cooper, Biederman, & Hummel, 1992; Riesenhuber & Poggio, 1999) and developmental psychologists (Landau, Smith, & Jones, 1988; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002) that human object recognition depends heavily on the shape of objects, more so than other features, such as colour, texture, size, etc. There could hardly be a more basic fact about human object recognition. As Hummel (2013) put it: "..the study of object recognition consist largely (although not exclusively) of the study of the mental representation of object shape, and the vast majority of theories of object recognition are, effectively, theories of the mental representation of shape". Accordingly, it might be expected that DNN models that perform well on predicting brain activations in visual cortex should also recognize objects largely based on shape.

However, in 2019, Geirhos et al. conducted a severe test of this hypothesis and showed that some of the same DNNs that do a good job in predicting brain activations in visual cortex exhibit a strong *texture-bias* rather than a shape-bias. In order to demonstrate this they presented DNNs with (a) photographs of images taken from ImageNet, (b) "texture" images that only included the texture of an object, and (c) and "style transfer" images in which the texture of one object was combined with the shape of another, as illustrated in Figure 4. The DNNs tended to classify the style transfer images on their texture rather than shape. In other words, DNNs trained on large image datasets were able to predict brain activations while relying on very different features of images compared to humans.

Figure 4

Style-transfer training stimuli from Geirhos et al. (2019) *An image from the ImageNet dataset (left) and 10 with the same shape/content but different texture/style (right).*

This Geirhos et al. (2019) study nicely highlights the importance of carrying out severe tests before drawing inferences about DNN-human similarities. This research also motivated future studies attempting to improve DNN-human correspondences with regards to shape bias, but again, strong conclusions have been drawn without severe testing. The first attempt was made by Geirhos et al. (2019) themselves, who used the style-transfer (Gatys, Ecker, & Bethge, 2016) to train DNNs to classify images. That is, DNNs were trained on image datasets where shape but not texture was diagnostic of category. Geirhos et al. (2019) found that DNNs trained in this way increased their shape-bias when classifying held-out style-transfer images. While this is an interesting machine learning solution to the problem as viewed from an engineering standpoint, there can be no doubt about its ecological (in)validity in terms of cognitive science. Not only do human infants not learn object recognition based on a set of labelled examples — a problem with all supervised learning models — they also do not learn based on examples where the texture of one category is superimposed on the shape of another category. This work inspired a related and more plausible solution by Hermann et al. (2020), who hypothesised that the texture-bias of DNNs may be due to the aggressive cropping of images for the sake of data augmentation during training. This cropping was thought to make texture more diagnostic than shape when classifying images. Indeed, Hermann et al. (2020) showed that decreasing the amount of cropping increased the shape-bias of DNNs. However, once again, no severe test was performed on whether the representations of shape or, indeed, the nature of shape-bias correspond to human shape-bias. Nevertheless, Hermann et al. (2020) write: "Our results indicate that apparent differences in the way humans and ImageNet-trained DNNs process images may arise not primarily from differences in their internal workings, but from differences in the data that they see" (Abstract). Much like the benchmark in Brain Score (Schrimpf et al., 2018), different models now compete on which one manages to show the most shape-bias on a style-transfer dataset. One of the leading models at the moment is a Vision Transformer with nearly 22 billion parameters, trained on a dataset of 4 billion images (Dehghani et al., 2023).

But showing that DNNs can be trained to classify style transfer images according to shape rather than texture is a weak test of the hypothesis that DNNs encode shape in a human-like way. Indeed, there are a wide variety of findings regarding how humans process shape for the sake of object identification, and current models fail to account for many of them (e.g., Baker & Elder, 2022; Baker, Lu, Erlikhman, & Kellman, 2018; German & Jacobs, 2020; Malhotra et al., 2023). Consider the study by Malhotra et al. (2022) who demonstrate that even when networks are trained to show shape-bias, the nature of this bias is different to humans in a critical way. The authors trained DNNs and humans to classify a set of novel objects that had both shape and non-shape features diagnostic of object category. Humans classified the novel objects based on their shapes and ignored highly predictive non-shape features. By contrast, DNNs did the opposite, and focused on the non-shape features. Critically, even when DNNs were pretrained trained to have a shape-bias (trained on the style transfer images), and even when almost all the weights were frozen (e.g., 49 out of 50 layers of ResNet50), the DNNs switched to learning based on the non-shape predictive feature of the novel objects. This result suggests that, unlike DNNs that show shape-bias, human shape-bias is not simply an artifact of learning the most predictive feature.

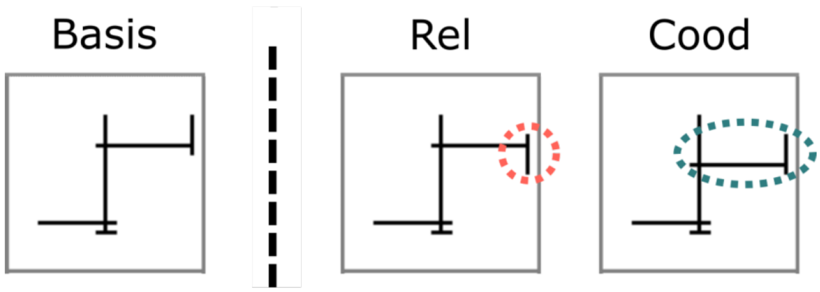

Figure 5

Example of an object and modified variants from Malhotra et al. (2023). *The basis object was modified to create two variants. (Rel) The first modification consisted of a categorical change of a relation between parts of the object. (Cood) The second modification preserved all relations but coordinates of some elements were shifted.*

In another study, Malhotra et al. (2023) go further and examine the nature of shape representations in DNNs that have a shape-bias and compare these to human shape representations. Humans have been shown to be sensitive to changes in relations between object parts (Stankiewicz & Hummel, 1996). Robust findings show that relation preserving changes often go unnoticed by human observers, while changes in relations between object parts are routinely noticed and interpreted as an important change either of the object or even the object category (Figure 5). In a series of simulations and experiments, Malhotra et al. (2023) tested DNNs (both standard and trained on the Stylized Images dataset) in order to determine whether DNN representations of shape share this property with humans. Performance measures as well as internal representations in this study indicated that DNNs do not share sensitivity to relational changes with humans. Malhotra et al. (2023) hypothesised that these differences between humans and DNNs originate from a difference in the goals of the two systems: while DNNs aim to classify their retinal images, humans aim to infer properties of distal objects that cause the retinal image.

We have focused on these two lines of research that have been particularly important with regards to claims regarding DNN-human similarities in the domain of vision, but this pattern of avoiding severe tests is widespread. For example, Zhou and Firestone (2019) claimed that there was a similarity between how humans and DNNs interpret adversarial images — i.e., nonsense images that were designed to fool the networks to confidently classify them. However, when this claim was rigorously tested by Dujmović, Malhotra, and Bowers (2020), it turned out that, for the vast majority of images and participants, there were significant differences in which these images were interpreted by DNNs and humans. Similarly, several researchers have posited that grid-cells — similar to those found in the entorhinal-hippocampal circuit — emerge as a result of training DNNs on path-integration (Banino et al., 2018; Cueva & Wei, 2018; Sorscher, Mel, Ganguli, & Ocko, 2019). However, when this claim was more severely tested by Schaeffer et al. (2022), they found that RNNs trained on path-integration almost never learn grid-like representations. Rather, the emergence of grid-like representations highly depends on a long list of specific decisions such as highly specific tuning of hyperparameters and design choices. Schaeffer et al. state: "...effectively baking in grid-cells into the task-trained networks. It is highly improbable that DL models of path integration would have produced grid cells as a novel prediction from task training, had grid cells not already been known to exist".

In some cases, the authors own findings do not support the conclusions they draw. For example, in the case of language, Schrimpf et al. (2021) report that transformer models predict nearly 100% of explainable variance in neural responses to written sentences and suggest that "a computationally adequate model of language processing in the brain may be closer than previously thought". However, the explainable variance is between 4-10% of the overall variance in three of the four datasets they analyze, and DNNs not only predict brain activation of language areas, but also non-language areas. Accordingly, it is not clear that these weak similarities have anything to do with language.

While severe testing of DNNs undermines many of the strong claims regarding DNN-human correspondences, it has not (yet) led to DNNs that do survive severe testing. Nevertheless, these studies provide critical insights into the nature of correspondence between

DNNs and humans and bring into focus broader issues around measuring similarity of representations between different systems. And most importantly, a better characterization of DNN-human similarities is a prerequisite for building better models of brains and minds.

How the peer-review process may contribute to the lack of severe testing

If severe testing has the potential to uncover critical insights about the relation between neural network models and human cognition, why is it frequently overlooked by the field? One of the reasons may be a bias against publishing *negative results* — that is, results highlighting dissimilarities between DNNs and humans.

It is certainly our impression that there are more published articles highlighting DNN-human similarities compared to differences. To see if this impression has any validity, we looked for articles published in three high-profile journals (PNAS, Nature Communications, and PLOS Computational Biology) from 2020 to present using a Google Scholar search that contained at least one of the following terms "DNN" or "DNN" or "DNNs" or "DNNs" as well as contained both "brain" and "object recognition" somewhere in the text. We then read the abstracts to confirm whether the papers were comparing DNNs to human vision (in some cases the articles returned from this search did not). Our judgements are somewhat subjective, and a few articles might be classified differently, but we expect there would be reasonable agreement in the following numbers: 15 hits in PNAS, with 10 out of 12 highlighting similarities, 26 hits in Nature Communications, with 10 out of 11 highlighting similarities, 29 hits in PLOS Computational Biology, with 14 of 16 highlighting similarities. See the Appendix where we go into these numbers in some more detail.

Of course, the observation that most published research highlights similarities rather than differences may have multiple causes. First, it may reflects the fact that DNNs are indeed similar to brains and that the published studies identify important similarities. However, this is unlikely, given (a) the numerous observations of differences in behaviour and internal representations highlighted by recent research (Bowers et al., 2022; Serre, 2019), (b) differences in architecture, learning algorithms, cost functions, learning environments, etc, and (c) the frequency with which conclusions are undermined by severe testing. Second, it is possible that researchers are excited about the promise of DNNs as models of brains given their phenomenal engineering successes and this biases researchers to focus on the similarities and ignore differences. Third, and relatedly, there may be a bias amongst reviewers and editors to publish results highlighting similarities and reject studies that highlight differences (similar to a bias of reporting significant effects and rejecting null results in psychology and many other disciplines; e.g., Simmons, Nelson, and Simonsohn (2011)). These latter two possibility may well interact: A bias to publishing "positive" results would likely incentivize researchers to look for DNN-human similarities and avoid severe testing that might make publishing more difficult.

In order to gain some insight into the possibility of a publication bias, we searched `openreview.net` and `neurips.cc`, which publish articles alongside openly accessible commentary from reviewers and editors for leading machine learning and AI conferences such as NeurIPS, ICML and ICLR. In reviewing these commentaries, we came across two types of objections that reviewers and editors frequently make in relation to studies empirically comparing DNNs and human cognition:

1. Reviewers feel that a negative result is not surprising as we already know that DNNs are not like humans. This type of comment places a premium on identifying results that are surprising over results that identify important differences between DNNs and human cognition. Here are some examples of this type of comment:

> **Example 1.1**: *"I find the overall conclusions unsurprising. It is to be expected that DNNs will perform quite poorly on data for which they were not trained. While a close comparison of the weakness of humans and DNNs would be very interesting, I feel the present paper does not include much analysis beyond the observation that new types of distortion break performance."* (Reviewer[a] comment on Geirhos et al. (2018))
>
> [a] Our intention here is not to pick on any particular reviewer but to reflect biases present in the field. Therefore, all examples chosen by us have anonymous reviewers.

> **Example 1.2:** *"...DNNs and human visual system are completely different systems, so it seems obvious at best to conclude that they may solve problems 'in a different manner' from each other."* (Reviewer comment on Malhotra et al. (2022))

> **Example 1.3**: *"In this empirical study, the authors attempt to identify a minimal entropy version of an image such that the image may be correctly classified by a human or computer... While identifying that humans are less sensitive to a reduction in resolution, this result is not terribly surprising given that networks are known to suffer from aliasing artifacts..."* (Reviewer comment on Carrasco, Hogan, and Pérez (2020)).

There are many other examples we could point to. For example, in their commentary on Bowers et al. (2022), Love and Mok (2023) write: "...we do not share [the authors'] enthusiasm for falsifying models that are a priori wrong and incomplete". Similarly, Tarr (in press) in his commentary, writes: "As a field we should have a productive discussion about what inferences we can draw from DNNs and other computational models (Guest and Martin, 2023). However, such discussions should involve less hyperbole... and less handwringing about what current models can't do; instead, they should focus on what DNNs can do".

It is difficult to know how frequent these types of comments are, but the fact that these comments exist at all shows that at least some reviewers see little value in reporting negative results while comparing DNNs and humans. And when negative results are published, the bar for getting these studies through the peer-review process seems to be higher. In Example 1.1, for example, the reviewer argues that it is *not* sufficient to show that DNN behaviour is different from humans, authors should also analyse *why* the behaviour differs. In contrast, we have many examples of positive results that have been reported in the literature (see for example Cadena et al., 2019; Cadieu et al., 2014; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Yamins et al., 2014; Zhuang et al., 2021) where studies report a correlation between DNN and a human / primate without identifying why this correlation exists.

In addition to the problems with incentivizing surprising results that we noted above, another problem with these comments is that they betray a lack of understanding of the value of negative results. Negative results do not just identify differences between DNNs and human cognition, they also frequently identify *how* the two systems differ. An investigation of this *how* question is non-trivial and, as we have argued in the previous section, has the potential to provide real insight into both human cognition and DNNs. By undervaluing such studies, the field risks ignoring key data points to guide future research. Fortunately, the Geirhos et al. (2019) study referred to in Example 1.1 has now been cited over 2000 times (according to Google Scholar) and provides a key constraint that guides existing results in developing DNNs better aligned to human visual system.

2. Reviewers feel that a study lacks novelty because it is an empirical study and does not suggest a new model that overcomes the observed dissimilarities. Here are some examples:

Example 2.1: *"[Authors] are only showing that the solution selected by the RNN does not follow the one that seems to be used by humans... [The] paper would really produce a more significant contribution [if] the authors can include some ideas about the ingredients of a RNN model, a variant of it, or a different type of model, must have to learn the compositional representation suggested by the authors."* (Reviewer comment on Lake and Baroni (2018))

Example 2.2: *"Overall, I think that the study can help to uncover systematic differences in visual generalization between humans and machines... The paper would have been much stronger if the first elements of algorithms that can counteract distortions were outlined. Although the empirical part is impressive and interesting, there was no theoretical contribution."* (Reviewer comment on Geirhos et al. (2018), NeurIPS)

> **Example 2.3**: Reviewer: *"This work demonstrates failures of relational networks on relational tasks, which is an important message. At the same time, no new architectures are presented to address these limitations."*
> Editor: *"While this paper does not propose solutions, it does present interesting "negative results" that should get some visibility in the workshop track."* (Editor & Reviewer comments on Kim, Ricci, and Serre (2018))

> **Example 2.4**: *"An elaborate human evaluation of two tasks, face identification and verification, has been conducted... AC agrees with the reviewers that albeit it's an important study, limited technical contribution (how to resolve existing model failures) and a narrow application domain (the paper studies face recognition and bias in face recognition) are two critical issues that place the contributions below the acceptance bar."* (Editor comment on Dooley et al. (2023))

Again, we have come across many other examples of this type of comment in our own work (see the following NeurIPS workshop talk by Bowers (2022) that provides multiple examples of reviewers and editors stating that falsification is not enough and that it is necessary to find "solutions" to make DNNs more like humans to publish: https://slideslive.com/38996707/researchers-comparing-dnns-to-brains-need-to-adopt-standard-methods-of-science.)

These comments again betray a clear preference for constructing a model—even a bad model—to a study that identifies an important limitation of existing models. In Example 2.3, for example, the paper is relegated to a workshop track because showing a critical failure of relational networks on relational tasks is deemed not worthy of the main conference. Publishing papers only if they report a new model creates a hurdle for reporting negative results. In view of these comments, it will not be surprising if many interesting observed differences between DNNs and humans go unreported.

A healthy back and forth within a field of research is to be expected. Indeed, if we look at the history of vision research, we will find opposing claims being tested by multiple research groups over years or even decades. Nuanced research, refining theories, severe testing – these are all necessary in order to push a field forward. However, the trend we described through examples above does not follow that healthy pattern. Rather, we see many examples of strong claims based on weak tests, while nuanced studies more severely testing these claims are under-represented in the literature. From the reviewer / editor comments we have highlighted above, it also seems clear that (at least some) reviewers do not view reporting negative results as valuable as constructing new models—a worrying trend for anyone interested in the benefits and limitations of using DNNs to understand human cognition.

Discussion

We make two general points in this paper that have a number of implications for the field of neuroAI. First, we highlight how the empirical research comparing DNNs to humans often fails to include severe testing of hypotheses, and this is leading to many unjustified conclusions. In our view, researchers need to modify their methods to include severe testing and consumers of research need to be more aware of these limitations when evaluating the research findings. Second, we consider why the field has largely avoided severe testing. Here we argue that the current review process is incentivising researchers to look for DNN-human similarities and downplay their differences. It will be important for reviewers and editors to evaluate the extent to which research includes severe testing of hypotheses in order to ensure claims regarding DNN-human similarities are well motivated.

With regards to the research, we have (i) elaborated on what such severe testing involves, and (ii) illustrated how the lack of severe testing characterises research comparing DNN and human vision in two separate lines of research. We could have focused on many other examples, and indeed, at the time of writing, there is much excitement regarding Large Language Models (LLMs), where we believe comparisons are being made with human cognition (Caucheteux, Gramfort, & King, 2022; Mahowald et al., 2023; Piantadosi, 2023; Schrimpf et al., 2021; Tuckute et al., 2023) without rigorously testing these claims. We simply focused on two lines of research in the domain of vision and object recognition that is closely related to our own work that illustrate the problems quite concretely.

It is important to be aware of the many different ways the lack of severe testing manifests itself. In some cases, severe tests have simply not been carried out and strong claims are made simply based on the observation of a correlation (see Bowers et al., 2022, for a number of examples). But in other cases, authors claim to have carried out strong tests of hypotheses but these tests fall short of the *severe tests* standard identified above. This happens in at least three forms. First, authors make a strong claim but, in reality, test a much weaker claim. For example, authors might claim that humans can decipher how DNNs classify adversarial images, but only test whether DNNs and humans agree in their classification of a small subset of these images under some limited experimental conditions. When the claims are tested more severely they are falsified (see Dujmović et al., 2020). Second, authors sometimes argue that their procedure represents a "strong test" that a model is similar to humans, but note in the Discussion or Appendix important qualifications that dramatically weaken the conclusions that should be drawn. For example, emphasizing in the body of the article that large language models account for 100% explainable variance of human BOLD signals, and noting in Appendix that explainable variance is extremely small and that similar BOLD prediction success occurs in non-language areas (Schrimpf et al., 2021). Third, authors may argue that an observed phenomenon emerges due to some feature of the training conditions, while in reality there are many other features of the training conditions (hyper-parameters, specific training dataset, etc.) that are required to observe the emergent phenomenon (Schaeffer et al., 2022). In each case, the authors (and readers) may fall prey to a kind of motte-and-bailey fallacy (Shackel, 2005), making a strong claim that is unwarranted by data and retreating to a more modest claim when challenged.

With regards to the incentives of the field that discourage severe testing, we argue that the current peer-review culture may be playing a role. Not only do most articles published in high profile journals make strong claims regarding DNN-human similarities, we provide examples of reviewers and editors undervaluing studies that challenge these conclusions through severe testing. Indeed, reviewers and editors often claim that "negative results" — i.e., results that falsify strong claims of similarity between humans and DNNs — are not enough and that "solutions" — i.e., models that report DNN-human similarities – are needed for publishing in the top venues (see example 2.1–2.4 quotes). Again, for many more examples, see Bowers et al. (2022).

Interestingly, similar issues have been raised in an engineering context in which there is no consideration of whether DNNs are like humans. In a NeurIPS talk, Kilian Weinberger (https://slideslive.com/38938218/the-importance-of-deconstructionpoints) criticizes the common practice of publishing models based on their performance without acting like a scientist and deconstructing the models to determine what aspects of the model are responsible for their success. He details three examples where his research team developed a complex model that solved an important task, but when they deconstructed the success of the model, it turned out that the key innovation was often trivial and not what they expected. Importantly, Weinberger highlights how the incentive structure in academia does not encourage this approach to research: before deconstruction, the paper was easily publishable, and after additional work that identifies the causal mechanisms of the success, the paper is more difficult to sell. Despite the obvious similarity to the situation with neuroAI, it is also important to emphasize an important difference. The main objective of the engineer is to solve a problem, and a complicated black box that solves an interesting problem may still be useful. By contrast, the main objective of researchers comparing DNNs to humans is to better understand the brain through DNNs. If apparent DNN-human similarities are mediated by qualitatively different systems, then the claim that DNNs are good models of brains is simply wrong.

More generally, there is now a widespread appreciation in many areas of science that a strong bias for publishing positive results (among other practices) is leading to a credibility crisis. Central to fixing this crisis is modifying the peer review process so that null results can be more easily published. Of course, the problem persists, but at least there is extensive discussion of the broader issues in the literature (e.g., see the special issue introduced by (Proulx & Morey, 2021), and concrete steps to better understand the problems and their root causes have been made (e.g., Buzbas, Devezer, & Baumgaertner, 2023; Devezer, Navarro, Vandekerckhove, & Buzbas, 2021; van Rooij & Baggio, 2021). Some solutions have been proposed, such as the Reproducibility Project: Psychology (https://osf.io/ezcuj/) where researchers attempt to replicate past findings (and where null results are commonplace), and the introduction of registered reports in some journals where manuscripts are accepted or rejected prior to carrying out the research to prevent a bias against negative outcomes, and multiple papers highlighting the problem. The specific solutions in psychology and other areas may not be appropriate to the current context, but there needs to be a similar recognition of the problems and active attempts to improve the processes by which papers are assessed. Of course, there is some recognition of these issues and some attempts to address the problems (e.g., the "I can't believe it's not better workshop" at NeurIPS that invites papers that report unexpected null findings or criticisms of standard practices), but the field is far behind others in this respect. Consequently, it is quite likely that many published claims regarding DNN-human similarities are false. We hope this article helps to fuel this conversation as it is needed for the development of better models of brains and mind that even the critics are hoping to see.

References

Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *Iscience*, *25*(9), 104913.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... others (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Bornet, A., Doerig, A., Herzog, M. H., Francis, G., & Van der Burg, E. (2021). Shrinking bouma's window: How to model crowding in dense displays. *PLoS computational biology*, *17*(7), e1009187.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... others (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74.

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, *10*(3), 221042.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, *10*(12), e1003963.

Carrasco, J., Hogan, A., & Pérez, J. (2020). *Laconic image classification: Human vs. machine performance.* Retrieved from https://openreview.net/forum?id=rJgPFgHFwr

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), 16327.

Cooper, E. E., Biederman, I., & Hummel, J. E. (1992). Metric invariance in object recognition: a review and further evidence. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *46*(2), 191.

Cueva, C. J., & Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International conference on learning representations.*

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., ... others (2023). Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09.*

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society open science*, *8*(3), 200805.

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., ... others (2022). The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*.

Dooley, S., Wei, G. Z., Downing, R., Shankar, N., Thymes, B. M., Thorkelsdottir, G. L., ... Goldstein, T. (2023). *Comparing human and machine bias in face recognition.* Retrieved from https://openreview.net/forum?id=wtQxtWC9bra

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2023). Obstacles to inferring mechanistic similarity using representational similarity analysis. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/05/01/2022.04.05.487135 doi: 10.1101/2022.04.05.487135

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *Elife*, *9*, e55978.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2414–2423).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=Bygh9j09KX

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, *31*.

German, J. S., & Jacobs, R. A. (2020). Can machine learning account for human visual object shape similarity judgments? *Vision Research*, *167*, 87–99.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*.

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, *118*(46), e2104779118.

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, *33*, 19000–19015.

Hummel, J. E. (2013). Object recognition. *Oxford handbook of cognitive psychology*, *810*, 32–46.

Jacob, G., Pramod, R., Katti, H., & Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, *12*(1), 1872.

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*(17), e2115302119.

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proceedings of the National Academy of Sciences*, *119*(27), e2115047119.

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

Kim, J., Ricci, M., & Serre, T. (2018). *Not-so-CLEVR: Visual relations strain feedforward neural networks.* Retrieved from https://openreview.net/forum?id=HymuJz-A-

Lake, B., & Baroni, M. (2018). *Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks.* Retrieved from https://openreview.net/forum?id=H18WqugAb

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Liu, J., Zhang, H., Yu, T., Ni, D., Ren, L., Yang, Q., ... others (2020). Stable maintenance of multiple representational formats in human visual short-term memory. *Proceedings of the National Academy of Sciences*, *117*(51), 32329–32339.

Love, B. C., & Mok, R. M. (2023, Mar). You can't play 20 questions with nature and win redux. Retrieved from psyarxiv.com/xaemv doi: 10.31234/osf.io/xaemv

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Malhotra, G., Dujmović, M., & Bowers, J. S. (2022). Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, *18*(5), e1009572.

Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. (2023). Human shape representations are not an emergent property of learning to classify objects. *Journal of Experimental Psychology: General*, *in press.*

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge ; New York, NY: Cambridge University Press.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118.

Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A., & Scherberger, H. (2020). A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proceedings of the national academy of sciences*, *117*(50), 32124–32135.

Piantadosi, S. (2023). Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz/007180*.

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, *16*(4), 671–681.

Rawski, J., & Baumont, L. (2022). Modern Language Models Refute Nothing.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences*, *118*(16), e2023123118.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in neural information processing systems*, *30*.

Saxena, R., Shobe, J. L., & McNaughton, B. L. (2022). Learning in deep neural networks and brains with similarity-weighted interleaved learning. *Proceedings of the National Academy of Sciences*, *119*(27), e2115229119.

Schaeffer, R., Khona, M., & Fiete, I. R. (2022). No Free Lunch from Deep Learning in Neuroscience: A Case Study through Models of the Entorhinal-Hippocampal Circuit. In *Advances in Neural Information Processing Systems.*

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., … others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, *5*, 399–426.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219.

Shackel, N. (2005). The vacuity of postmodernist methodology. *Metaphilosophy*, *36*(3), 295–320.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological science*, *13*(1), 13–19.

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, *119*(43), e2200800119.

Sorscher, B., Mel, G., Ganguli, S., & Ocko, S. (2019). A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems*, *32*.

Stankiewicz, B. J., & Hummel, J. E. (1996). Categorical relations in shape perception. *Spatial vision*, *10*(3), 201–236.

Tarr, M. J. (in press). My pet pig won't fly and i want a refund. *Behavioral and Brain Sciences*, commentary.

Tsao, T., & Tsao, D. Y. (2022). A topological solution to object segmentation and tracking. *Proceedings of the National Academy of Sciences*, *119*(41), e2204248119.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., . . . Fedorenko, E. (2023). Driving and suppressing the human language network using large language models. *bioRxiv*.

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, 682–697.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, *12*(1), 2065.

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*, *111*(23), 8619–8624.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., . . . others (2023). Catalyzing next-generation artificial intelligence through neuroai. *Nature Communications*, *14*(1), 1597.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature communications*, *10*(1), 1334.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.

Appendix

In Google Scholar we used the search terms (1) “DNN” or “DNN” or “DNNs” or “DNNs”; (2) “brain” and "object recognition"; and (3) a specific journal or conference proceeding. We then read the abstract to assess whether indeed the paper was assessing the similarity of a DNN to human (or monkey) vision. In the case of searching the journal Proceedings of the National Academy of Sciences we obtained 14 hits.

1. Mehrer et al. (2021) - An ecologically motivated image dataset for deep learning yields better models of human vision.

2. Golan et al. (2020) - Controversial stimuli: Pitting neural networks against each other as models of human cognition.

3. Sorscher, Ganguli, and Sompolinsky (2022) - The neural architecture of language: Integrative modeling converges on predictive processing.

4. Firestone (2020) - Performance vs. competence in human–machine comparisons.

5. Sablé-Meyer et al. (2021) - Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity.

6. Schrimpf et al. (2021) - The neural architecture of language: Integrative modeling converges on predictive processing.

7. Zhuang et al. (2021) - Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences.

8. Hannagan, Agrawal, Cohen, and Dehaene (2021) - Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading.

9. Michaels, Schaffelhofer, Agudelo-Toro, and Scherberger (2020) - A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping.

10. Saxena, Shobe, and McNaughton (2022) - Learning in deep neural networks and brains with similarity-weighted interleaved learning.

11. Jozwik et al. (2022) - Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models.

12. Jagadeesh and Gardner (2022) - Texture-like representation of objects in human visual cortex.

13. Liu et al. (2020) - Stable maintenance of multiple representational formats in human visual short-term memory.

14. Tsao and Tsao (2022) - A topological solution to object segmentation and tracking.

Articles 13 and 14 can be excluded as they are not addressing the relation between DNNs and human vision. Of the 12 remaining relevant studies, all emphasize the similarities of DNNs and human vision or the promise of DNNs as models of human vision, with the partial exception of articles 2 and 5. Article 2 highlights the value of designing a new type of stimulus (controversial stimuli) that provide a more severe tests of DNN-human vision correspondences (much in line with the approach adopted here). The authors reported lower RSA scores for models tested with these images. Article 5 shows that human vision is sensitive the geometric shape regularities whereas baboon vision and feed-forward DNNs are not. The authors suggest that symbolic processes may be missing from current DNNs.

More briefly, a similar outcome was obtained when we used the same search terms for Nature Communications, with 29 hits, and after reading the abstracts we identified 11 papers that assess the similarity of DNNs and human vision, with 10 papers emphasizing similarities. The one clear exception highlights how RSA scores are much smaller than past reports with a new fMRI dataset:

- Xu and Vaziri-Pashkam (2021) - Limits to visual representational correspondence between convolutional neural networks and the human brain.

Adopting a somewhat looser criterion you might note that the article by Jacob, Pramod, Katti, and Arun (2021). also highlighted some limitations of DNNs as models of vision:

- Jacob et al. (2021) - Qualitative similarities and differences in visual object representations between brains and deep networks.

But the later authors are clearly highlighting the promise of DNNs, concluding the abstract with: "These findings indicate sufficient conditions for the emergence of these phenomena in brains and deep networks, and offer clues to the properties that could be incorporated to improve deep networks".

Similarly, using the same search terms, we obtained 30 hits in PLOS Computational Biology and estimate that 14 out of 16 studies highlight the promise of DNNs as models of human vision, the two exceptions being:

- Malhotra et al. (2022) - Feature blindness: a challenge for understanding and modelling visual object recognition.
- Bornet, Doerig, Herzog, Francis, and Van der Burg (2021) - Shrinking Bouma's window: How to model crowding in dense displays.

The first article highlights how current DNNs do not have the same inductive biases to rely on shape when learning to classify novel stimuli. The second article shows that DNNs cannot account for the phenomena of "uncrowding", although they did find some non-DNN models could, including Capsule networks (Sabour, Frosst, & Hinton, 2017).